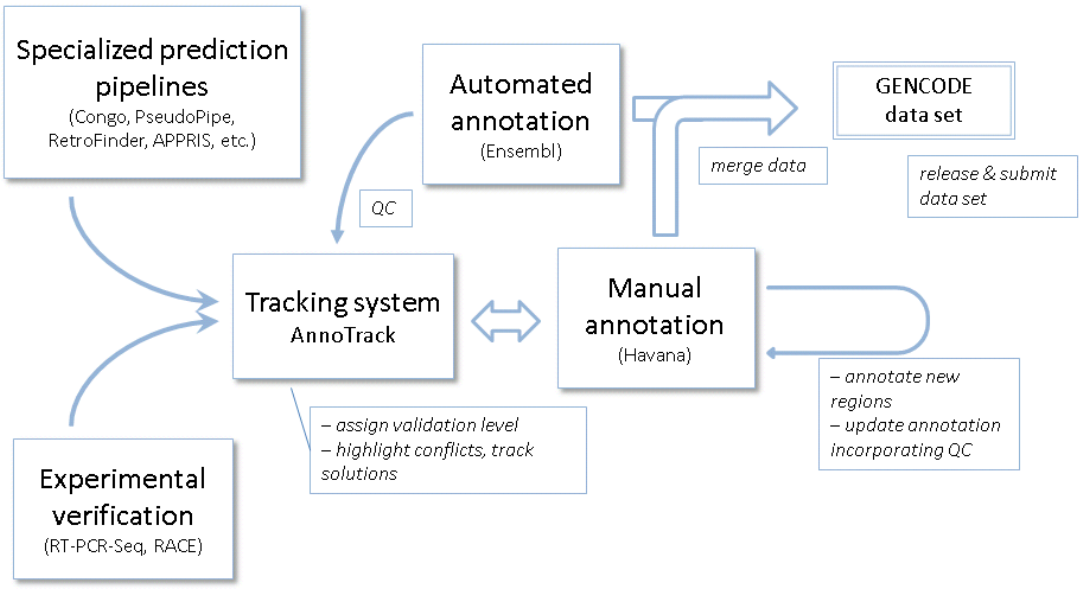
Overview of the GENCODE reference gene set

Aim

This module will give an overview of the GENCODE gene set that is available from the genome browsers and explain how ENCODE data is integrated to improve the set.

Introduction

Schematic showing interconnection between different GENCODE pipelines



HAVANA (Human and Vertebrate Analysis and Annotation) group at the Wellcome Trust Sanger Institute (WTSI) perform manual genome annotation. Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases (including cross-species) and a series of *ab initio* gene predictions. Annotation is based on supporting evidence, which is external sequence such as ESTs, cDNAs and protein. There are multiple biotypes that reflect confidence levels and there are additional data sources included as DAS tracks (e.g. CAGE tags, RNAseq).

The GENCODE reference gene set (http://www.gencodegenes.org/) is produced in collaboration between HAVANA and Ensembl and is available for human and mouse. The HAVANA group provides manual annotation of the

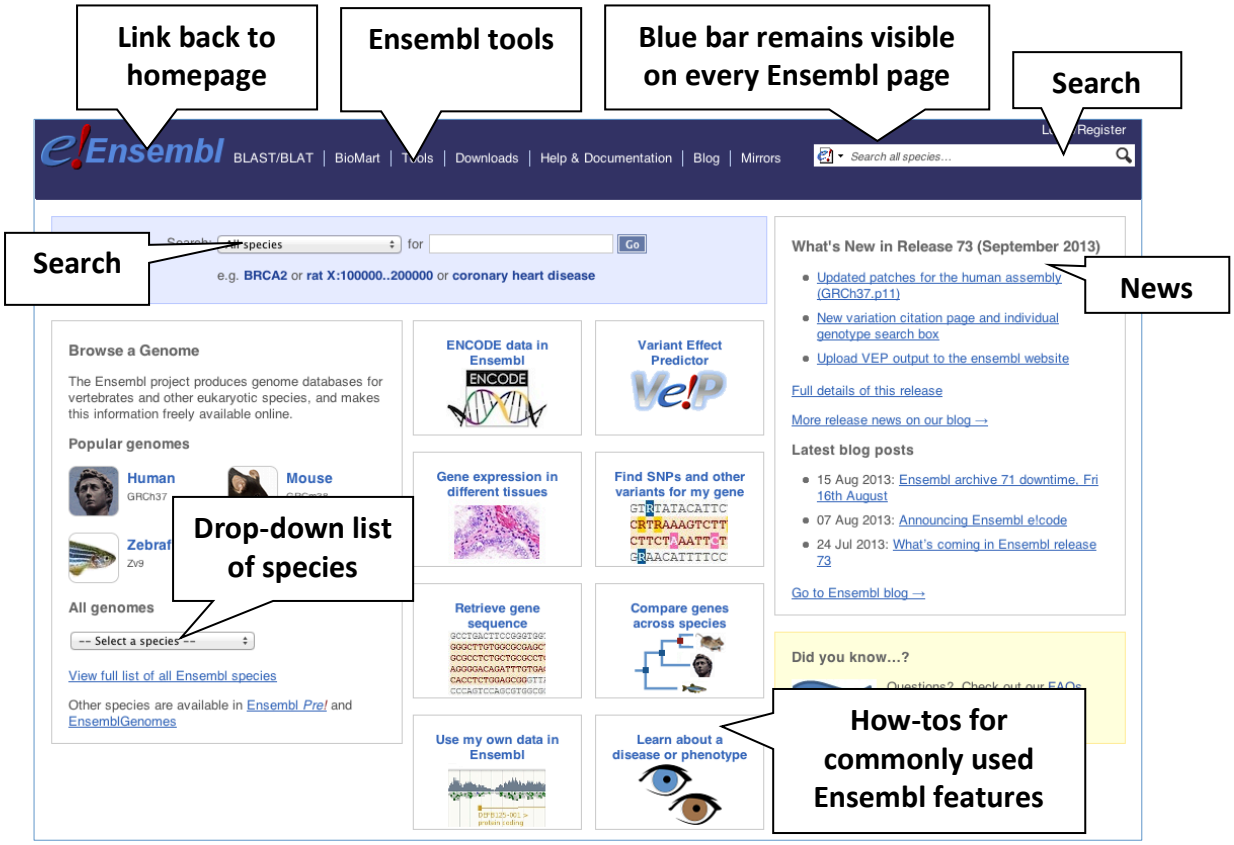
genes and transcripts onto the genome, whilst the Ensembl group performs automatic gene annotation. These data are merged to produce the GENCODE gene set, which is used by a variety of projects, such as Ensembl, ENCODE, 1000 genomes, UCSC and more.

The complete GENCODE set is available to view in genome browsers, such as Ensembl (http://www.ensembl.org/) and UCSC (http://genome.ucsc.edu/). The manually annotated genes only can be seen in the Vega (Vertebrate Genome Annotation) database (http://vega.sanger.ac.uk/).

The GENCODE gene set is an important contributor to the Consensus CDS (CCDS) project, which is a collaborative effort between the European Bioinformatics Institute (EBI), the National Centre for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI) and the University of California at Santa Cruz (UCSC). The aim of the project is to identify a core set of human protein coding regions that are consistently annotated between the different institutes. The long-term goal is to support convergence towards a standard set of gene annotations. The CCDS gene set is generated by Ensembl and NCBI and there is extensive QC by WTSI, NCBI and UCSC. A set of guidelines have been developed for the annotation of coding sequence regions by the collaborating Institutes, and any changes to the CCDS set have to be agreed by all three sites.

Demo: Looking at GENCODE genes in genome browsers

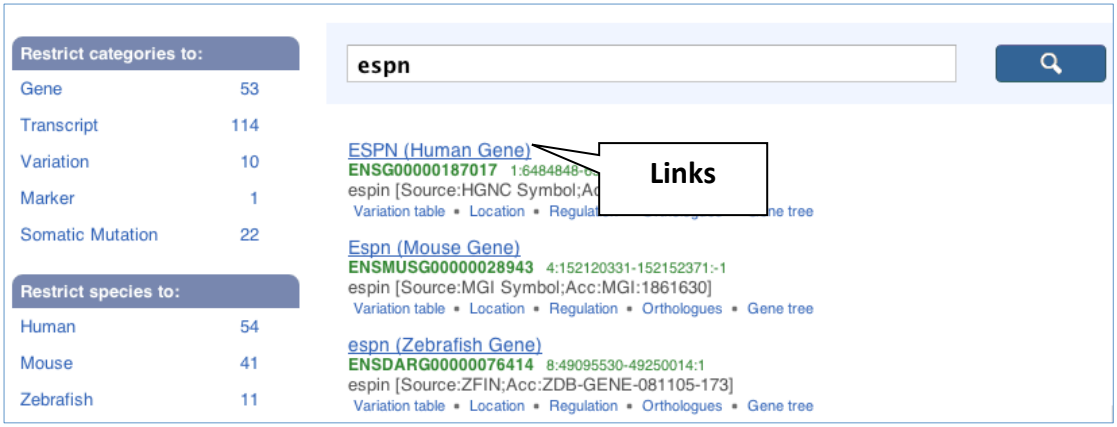
The front page of Ensembl is found at ensembl.org. It contains lots of information and links to help you navigate Ensembl:



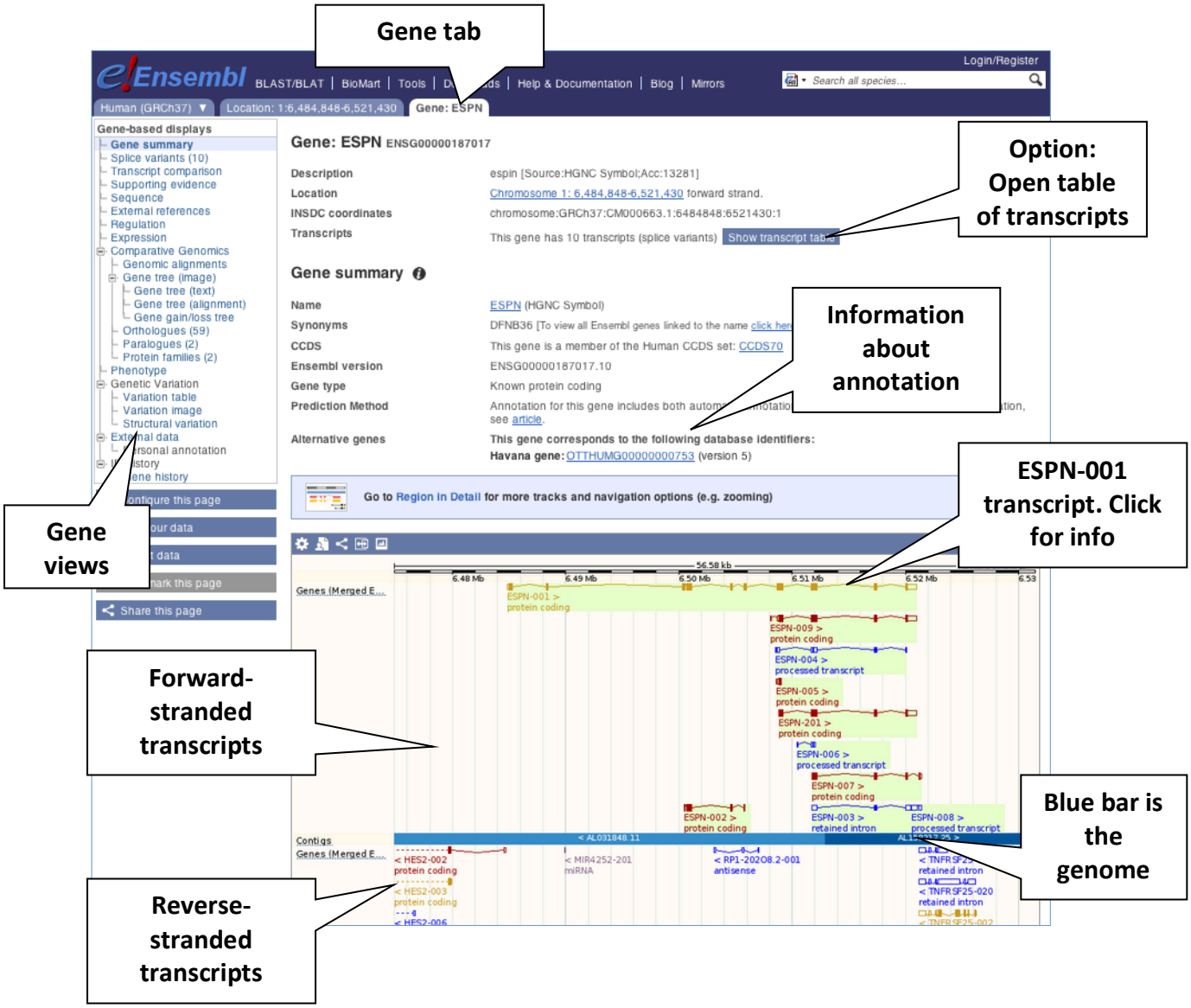
We're going to look at the human *ESPN* gene. This gene encodes a multifunctional actin-bundling protein with a major role in mediating sensory transduction in various mechanosensory and chemosensory cells. Mutations in this gene are associated with deafness (http://tinyurl.com/espn-ncbi-gene).

From ensembl.org, type *ESPN* 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.

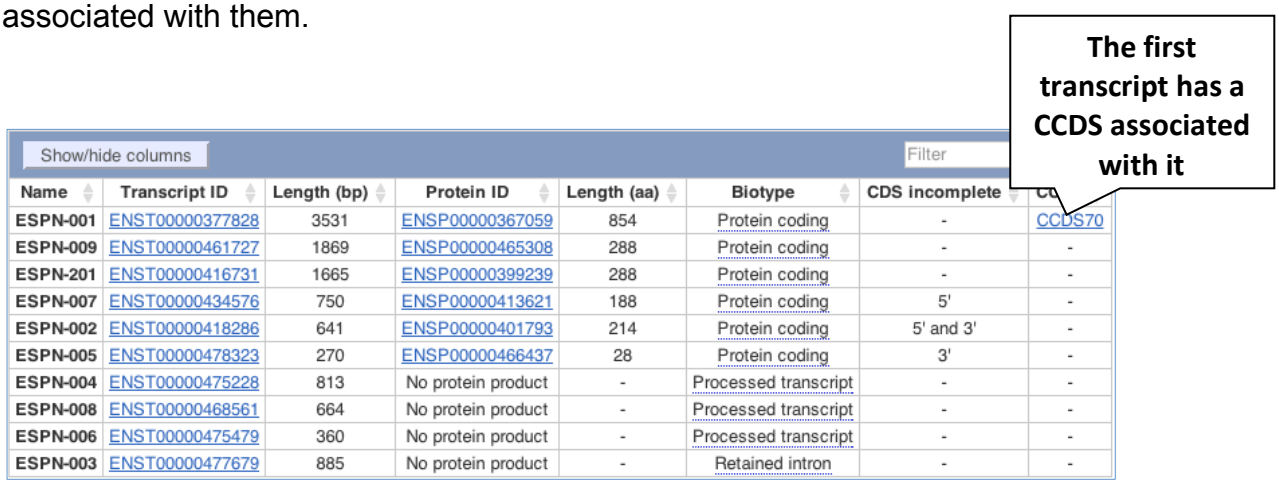


Click on the gene name or Ensembl ID. The **Gene tab** should open:



From this page we can see that the gene was annotated by both Ensembl automatic and Havana manual annotation. It is also a member of the CCDS

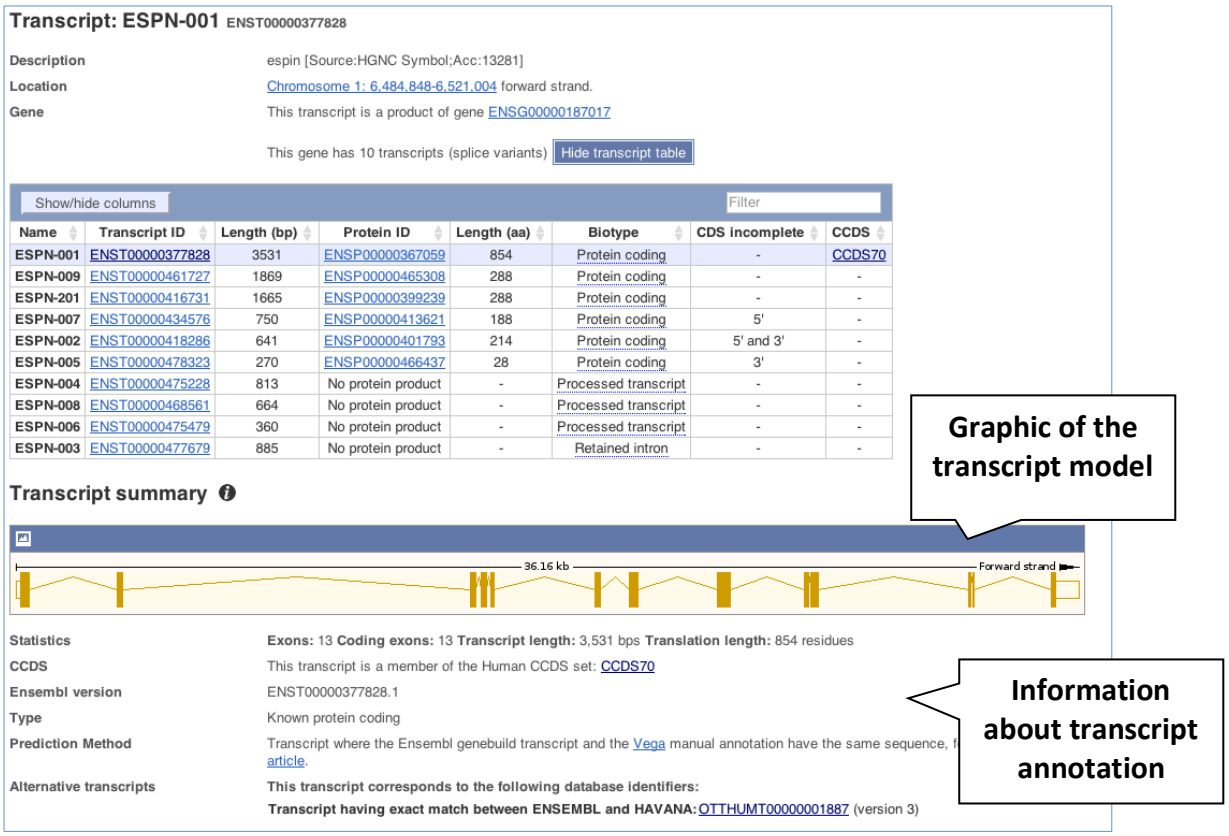
set. Click on Show transcript table to see which transcripts have a CCDS associated with them.



Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
ESPN-001	ENST00000377828	3531	ENSP00000367059	854	Protein coding	-	CCDS70
ESPN-009	ENST00000461727	1869	ENSP00000465308	288	Protein coding	-	-
ESPN-201	ENST00000416731	1665	ENSP00000399239	288	Protein coding	-	-
ESPN-007	ENST00000434576	750	ENSP00000413621	188	Protein coding	5'	-
ESPN-002	ENST00000418286	641	ENSP00000401793	214	Protein coding	5' and 3'	-
ESPN-005	ENST00000478323	270	ENSP00000466437	28	Protein coding	3'	-
ESPN-004	ENST00000475228	813	No protein product	-	Processed transcript	-	-
ESPN-008	ENST00000468561	664	No protein product	-	Processed transcript	-	-
ESPN-006	ENST00000475479	360	No protein product	-	Processed transcript	-	-
ESPN-003	ENST00000477679	885	No protein product	-	Retained intron	-	-

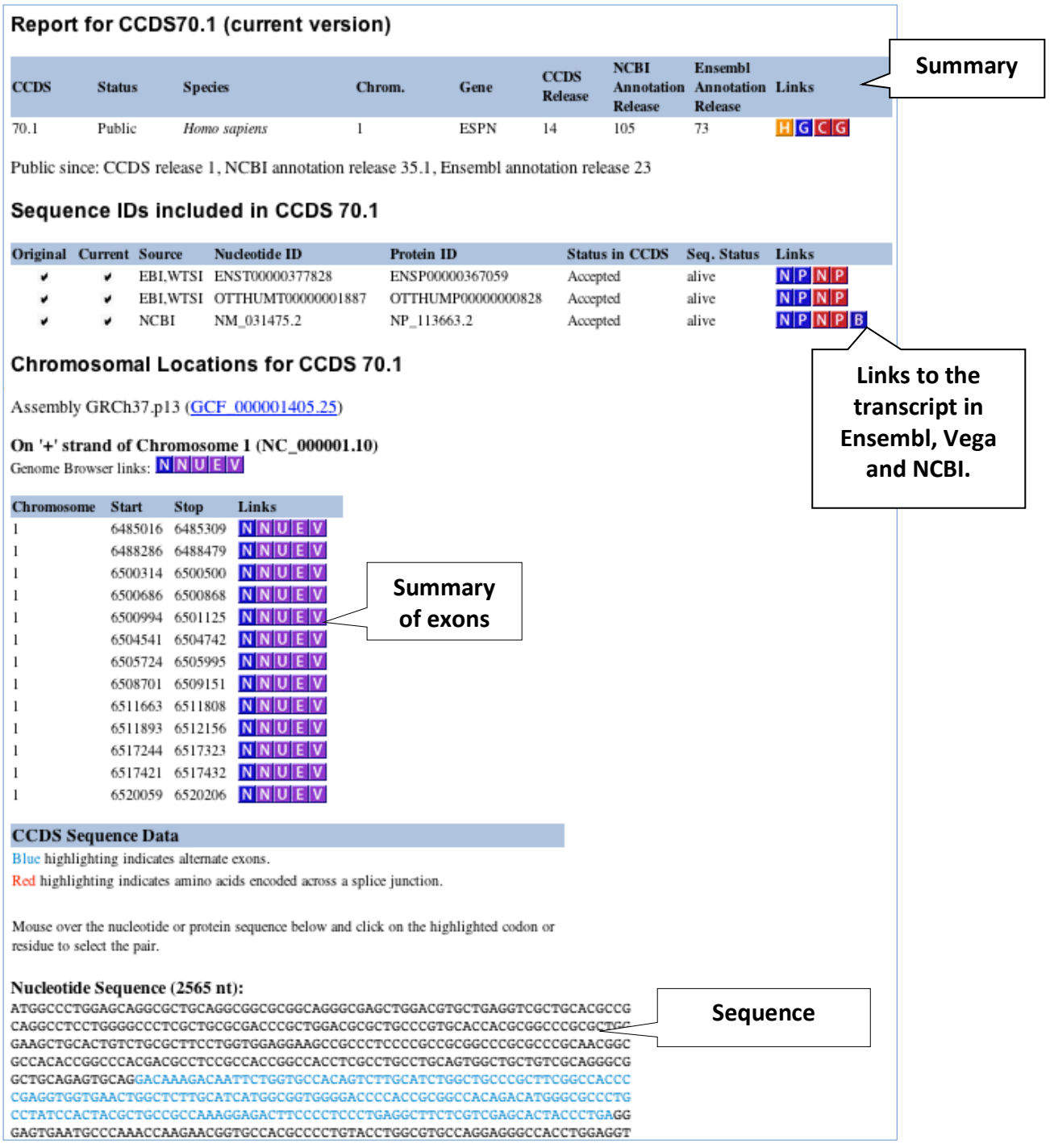
Select the top transcript ESPN-001 to go to the transcript tab.

This is the transcript associated with the CCDS transcript. It is shown in gold because it has identical annotation from the Ensembl automatic and Havana manual annotation.



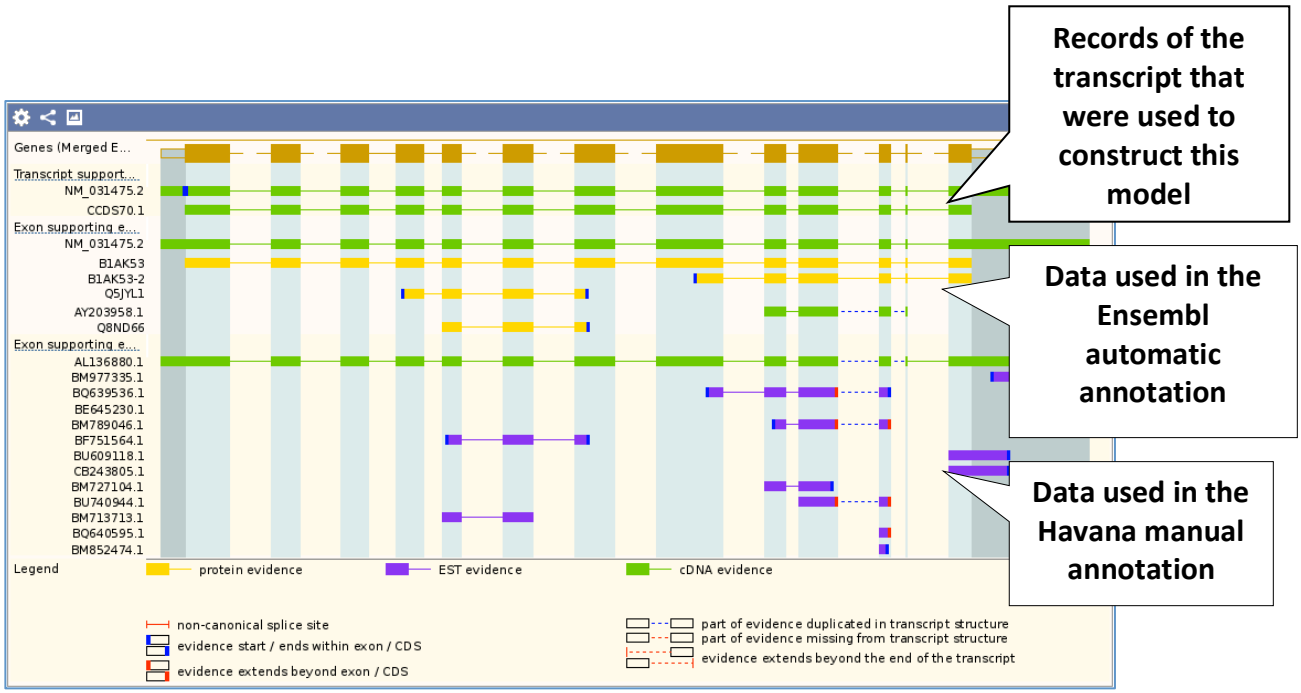
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
ESPN-001	ENST00000377828	3531	ENSP00000367059	854	Protein coding	-	CCDS70
ESPN-009	ENST00000461727	1869	ENSP00000465308	288	Protein coding	-	-
ESPN-201	ENST00000416731	1665	ENSP00000399239	288	Protein coding	-	-
ESPN-007	ENST00000434576	750	ENSP00000413621	188	Protein coding	5'	-
ESPN-002	ENST00000418286	641	ENSP00000401793	214	Protein coding	5' and 3'	-
ESPN-005	ENST00000478323	270	ENSP00000466437	28	Protein coding	3'	-
ESPN-004	ENST00000475228	813	No protein product	-	Processed transcript	-	-
ESPN-008	ENST00000468561	664	No protein product	-	Processed transcript	-	-
ESPN-006	ENST00000475479	360	No protein product	-	Processed transcript	-	-
ESPN-003	ENST00000477679	885	No protein product	-	Retained intron	-	-

Click on the CCDS370 to go to the CCDS record. This will open in a new tab in your browser.



This page summarises the CCDS transcript.

Go back to the Ensembl page and click on Supporting evidence in the left-hand menu.



You can see that Ensembl and Havana used different pieces of evidence to construct their transcript model, yet still came up with the same model, demonstrating how reliable the model is.

Click on General Identifiers in the left-hand menu. This lists records of the transcript and its protein product in other databases.

General identifiers

This transcript corresponds to the following database identifiers:

Show All entries Filter

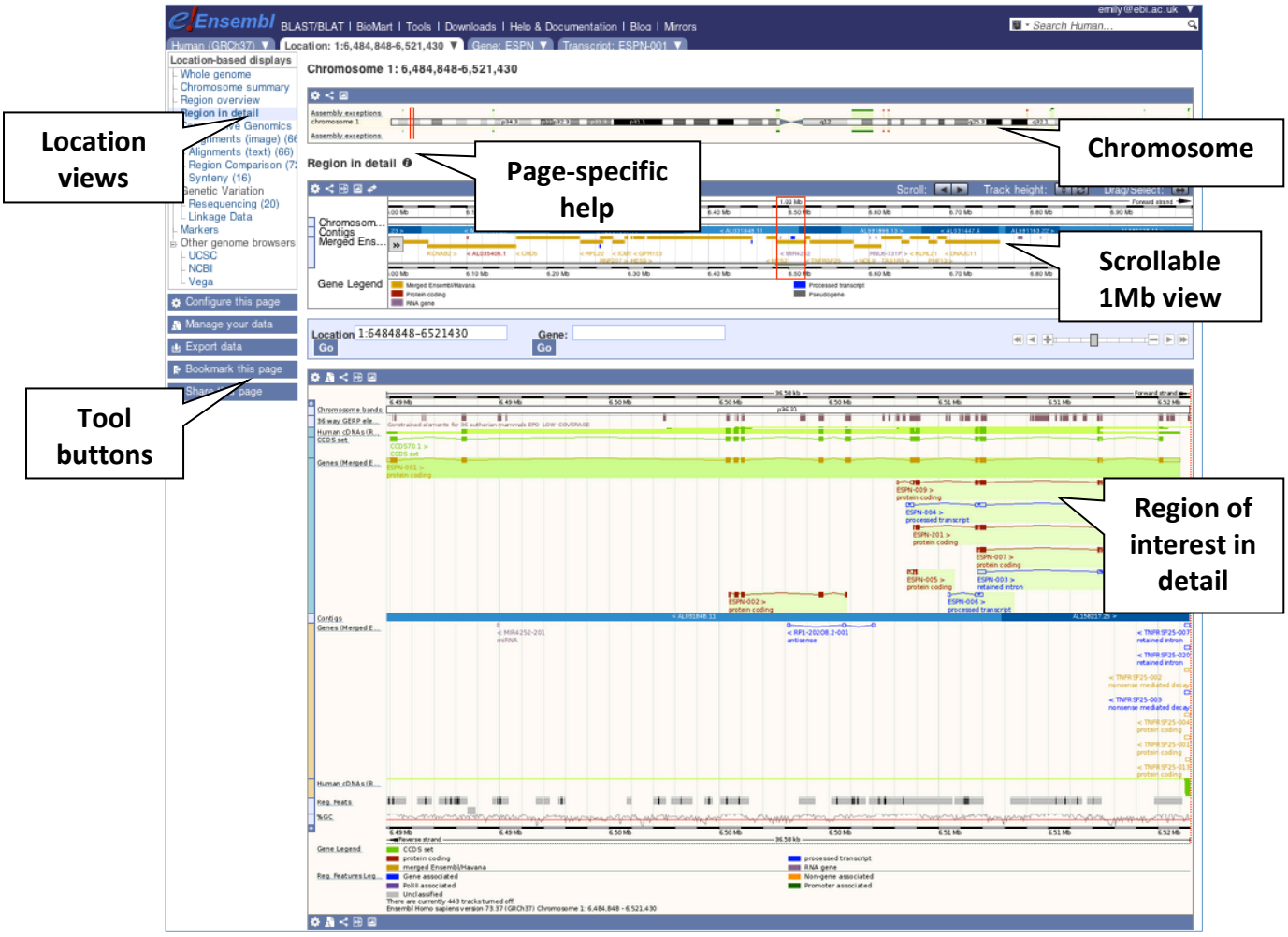
External database	Database identifier
HGNC Symbol	ESPN espin [view all locations]
UniParc	UPI000013D2B6 [view all locations]
CCDS	CCDS70.1 [view all locations]
UniProtKB/Swiss-Prot	ESPN_HUMAN [align] Espin [view all locations]
RefSeq peptide	NP_113663.2 [Target %id: 100; Query %id: 100] [align] espin [view all locations]
RefSeq mRNA	NM_031475.2 [align] [view all locations]
UCSC Stable ID	uc001amy.3 [view all locations]
Human Protein Atlas	HPA028674 [view all locations] HPA028674 [view all locations]
European Nucleotide Archive	AF134401 [align] [view all locations] AL031848 [align] [view all locations] AL136880 [align] [view all locations] AL158217 [align] [view all locations] AY203958 [align] [view all locations] CH471130 [align] [view all locations]
HGNC transcript name	ESPN-001 espin [view all locations]
INSDC protein ID	AAD24480.1 [align] [view all locations] AAP34481.1 [align] [view all locations] CAB66814.1 [align] [view all locations] CAI19773.1 [align] [view all locations] CAI22163.1 [align] [view all locations] EAW71537.1 [align] [view all locations]

We can also see genes and transcripts in a location. Click on the tab saying Location 1:6,484,848-6,521,530 at the top of the page.

Click on the button  to view page-specific help.

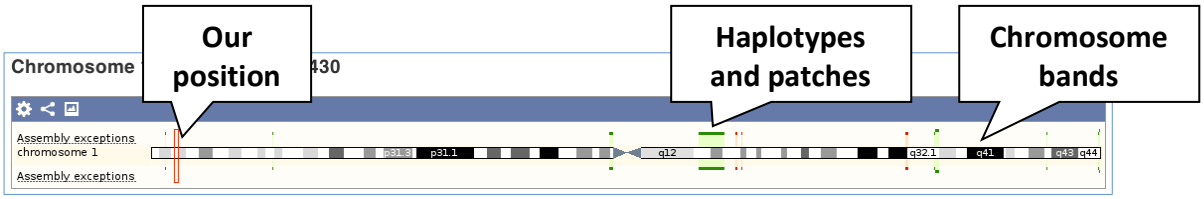
The help pages provide links to Frequently Asked Questions, a Glossary, Video Tutorials, and a form to Contact HelpDesk.

There is a help video on this page at http://youtu.be/tKEvgPUq94.

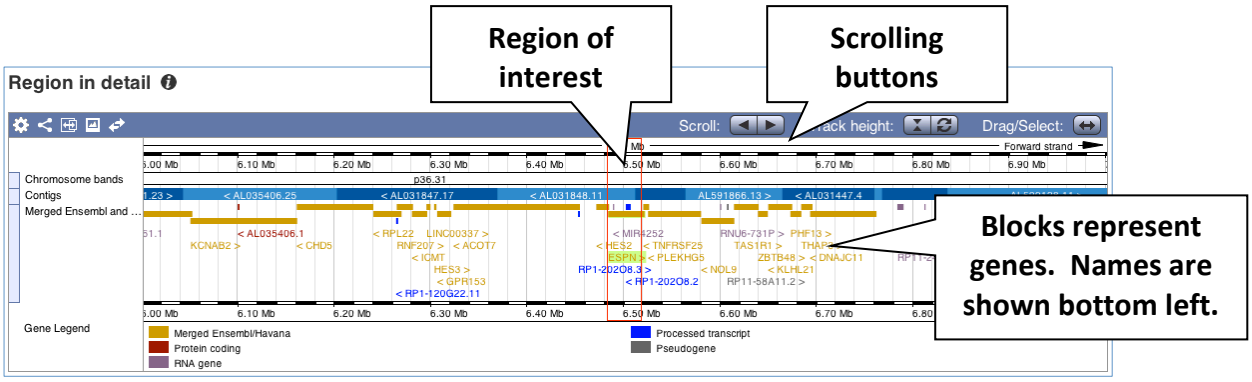


The Region in detail page is made up of three images, let's look at each one on detail.

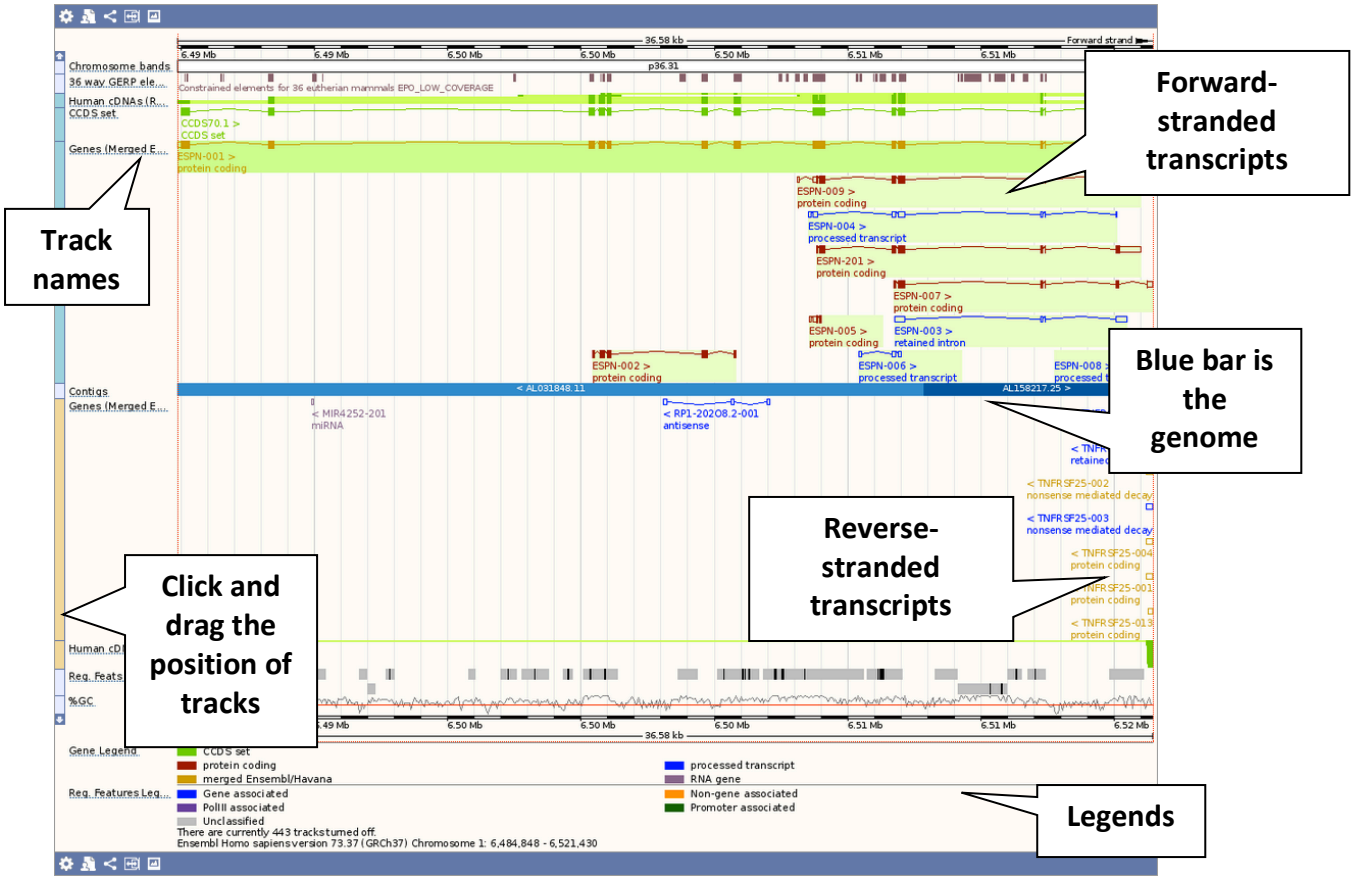
The first image shows the chromosome:



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

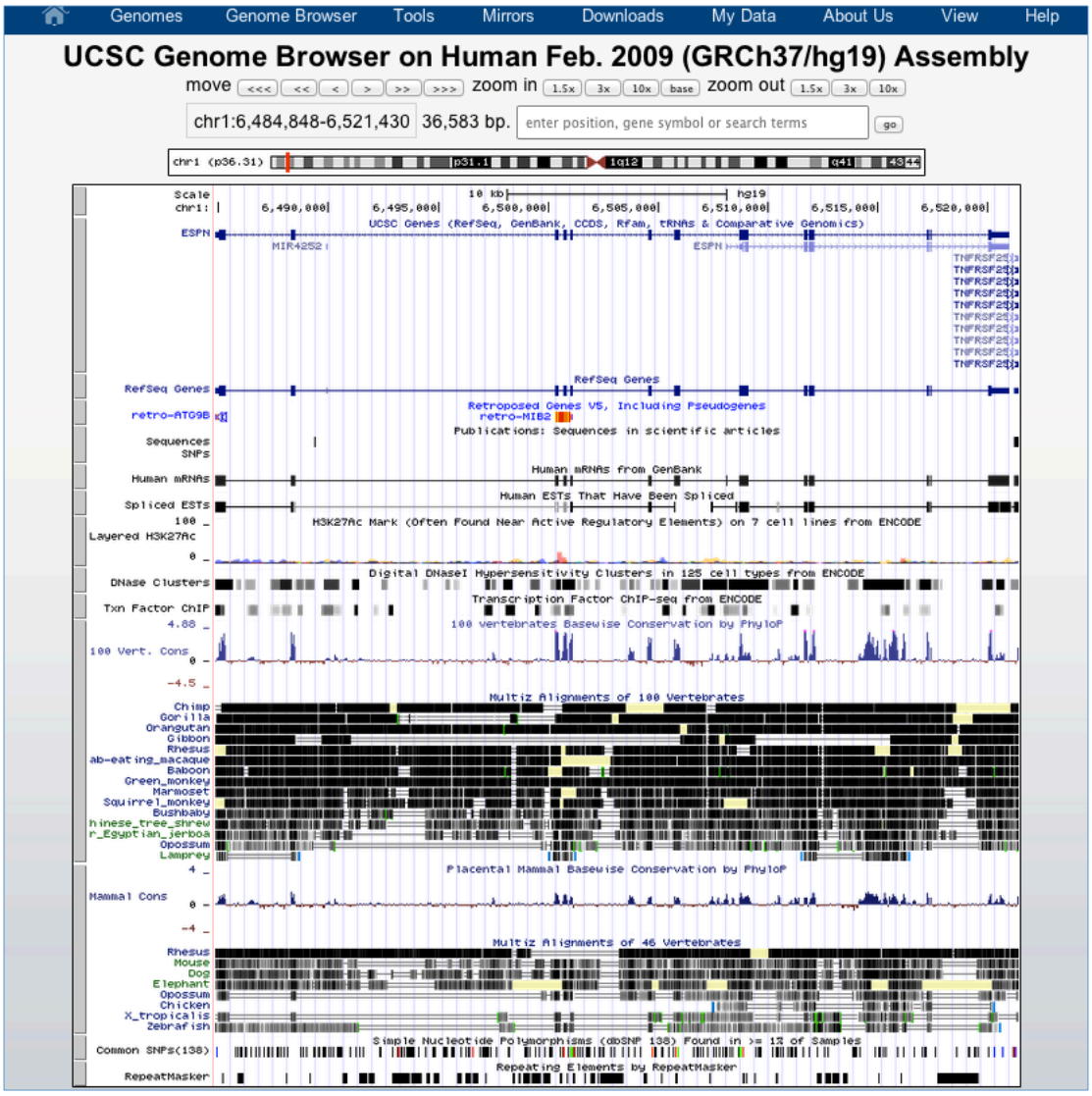


The third image is a detailed, configurable view of the region.

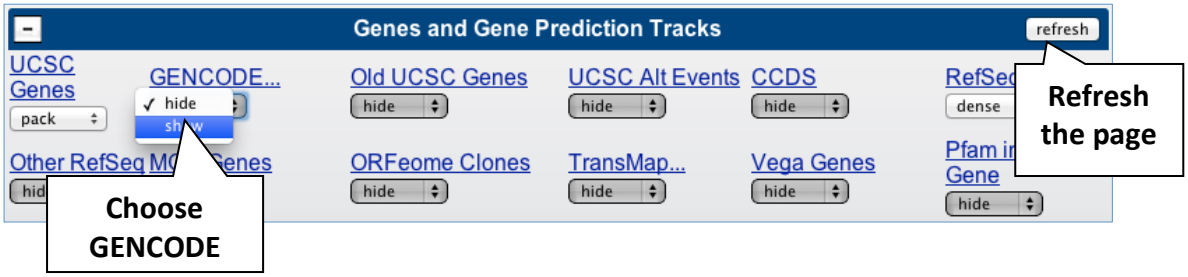


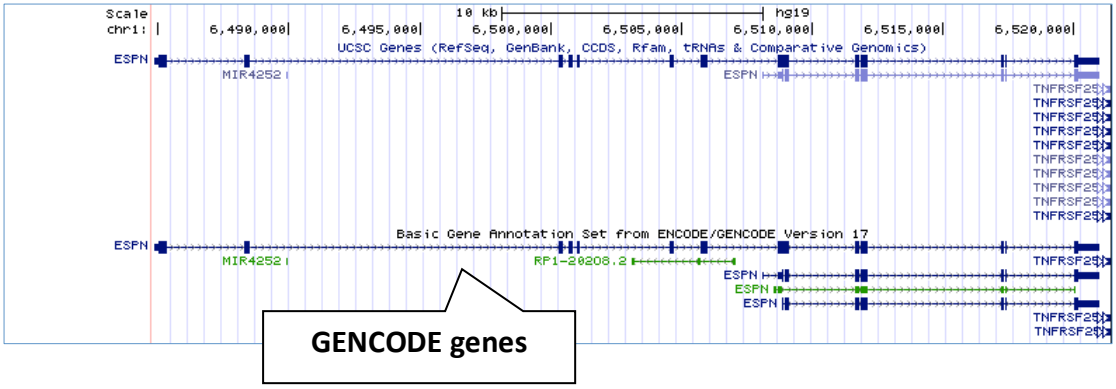
We can see the GENCODE genes in this view, with the CCDS plotted alongside (since CCDS transcripts lack UTRs).

Click on UCSC in the left-hand menu to see the same region in UCSC. This will open in a new tab.



The GENCODE gene set is not shown by default. Scroll down to Genes and Prediction tracks then select the drop down under GENCODE and chose show, then click on refresh.





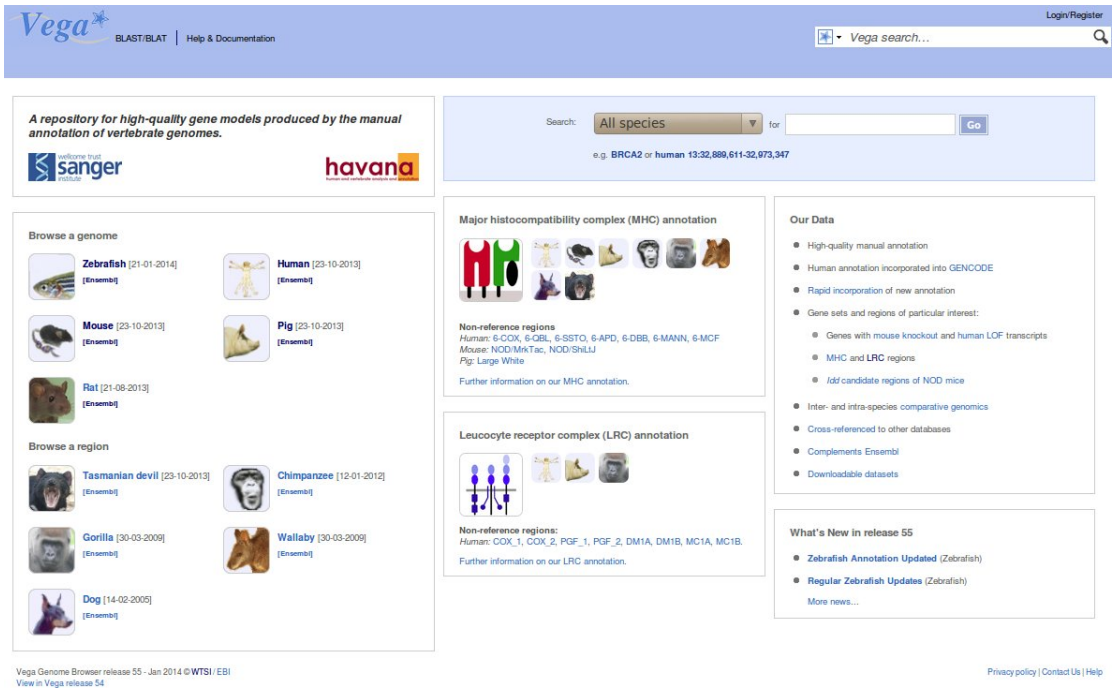
Click on the transcripts to see information about them, including links to Ensembl and Vega.

GENCODE Transcript Annotation ENST00000377828.1 (ESPN)

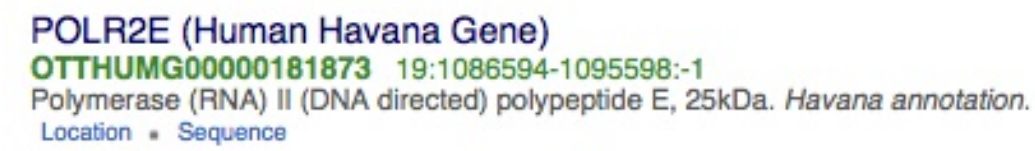
	Transcript	Gene
Gencode id	ENST00000377828.1	ENSG00000187017.10
HAVANA manual id	OTTHUMT00000001887.3	OTTHUMG00000000753.5
Position	chr1:6484848-6521004	chr1:6484848-6521004
Strand	+	
Biotype	protein_coding	protein_coding
Status	KNOWN	KNOWN
Annotation Level	manual (2)	
Annotation Method	manual & automatic	manual & automatic
Transcription Support Level	tsl1	
HGNC gene symbol	ESPN	
CCDS	CCDS70.1	
GeneCards	ESPN	
APPRIS	ENST00000377828.1	ENSG00000187017.10

Demo: Vega update genes

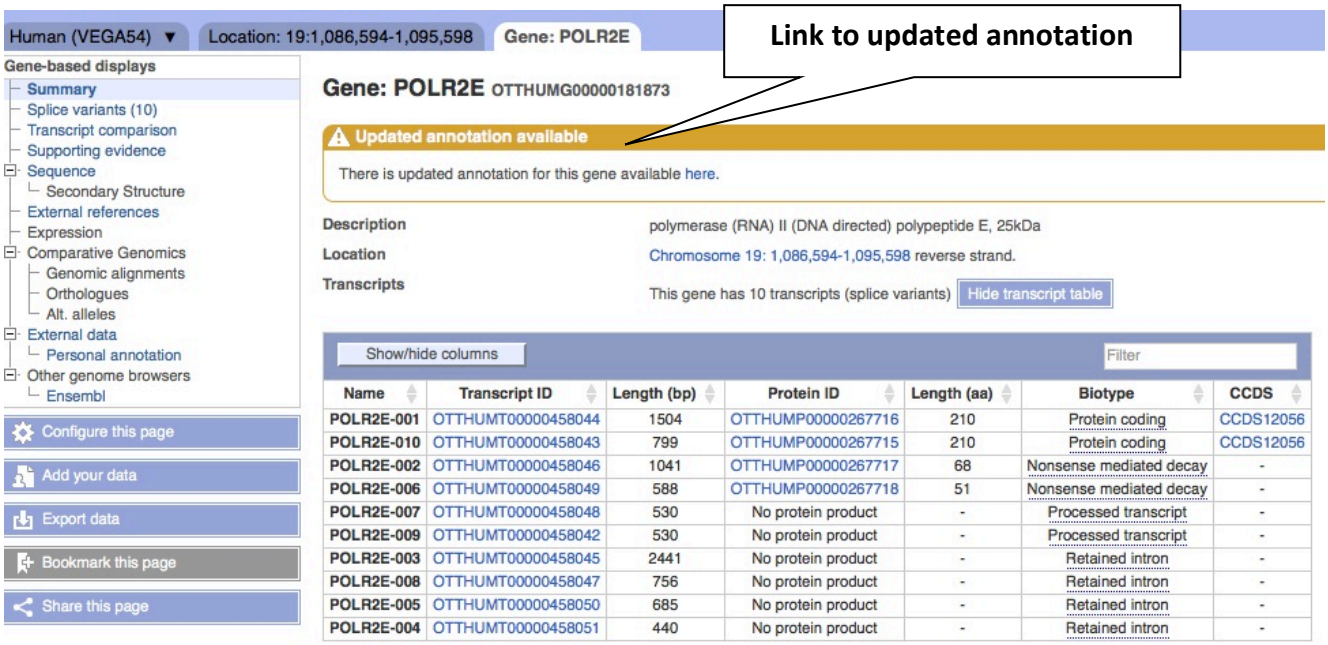
Start at the homepage for Vega (http://vega.sanger.ac.uk).



Click on human and search for the POLR2E gene.



Select the top result.



Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
POLR2E-001	OTTHUMT00000458044	1504	OTTHUMP00000267716	210	Protein coding	CCDS12056
POLR2E-010	OTTHUMT00000458043	799	OTTHUMP00000267715	210	Protein coding	CCDS12056
POLR2E-002	OTTHUMT00000458046	1041	OTTHUMP00000267717	68	Nonsense mediated decay	-
POLR2E-006	OTTHUMT00000458049	588	OTTHUMP00000267718	51	Nonsense mediated decay	-
POLR2E-007	OTTHUMT00000458048	530	No protein product	-	Processed transcript	-
POLR2E-009	OTTHUMT00000458042	530	No protein product	-	Processed transcript	-
POLR2E-003	OTTHUMT00000458045	2441	No protein product	-	Retained intron	-
POLR2E-008	OTTHUMT00000458047	756	No protein product	-	Retained intron	-
POLR2E-005	OTTHUMT00000458050	685	No protein product	-	Retained intron	-
POLR2E-004	OTTHUMT00000458051	440	No protein product	-	Retained intron	-

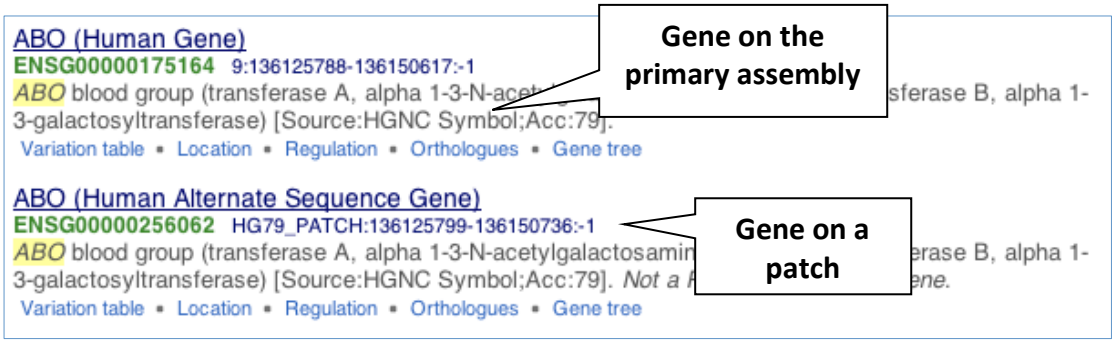
The current version of Vega has 10 splice variants for POLR2E, but there is updated annotation available. Click on the update link.

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
POLR2E-001	OTTHUMT00000458044	2831	OTTHUMP00000267716	210	Protein coding	-	-
POLR2E-013	OTTHUMT00000473950	1749	OTTHUMP00000274778	210	Protein coding	-	-
POLR2E-010	OTTHUMT00000458043	1096	OTTHUMP00000267715	210	Protein coding	-	-
POLR2E-009	OTTHUMT00000474120	932	OTTHUMP00000274865	184	Protein coding	-	-
POLR2E-014	OTTHUMT00000474115	487	OTTHUMP00000274861	134	Protein coding	5'	-
POLR2E-012	OTTHUMT00000473949	1286	OTTHUMP00000274777	204	Nonsense mediated decay	-	-
POLR2E-002	OTTHUMT00000458046	1214	OTTHUMP00000267717	68	Nonsense mediated decay	-	-
POLR2E-006	OTTHUMT00000458049	1105	OTTHUMP00000267718	51	Nonsense mediated decay	-	-
POLR2E-007	OTTHUMT00000458048	1052	OTTHUMP00000274864	83	Nonsense mediated decay	5'	-
POLR2E-011	OTTHUMT00000474116	589	No protein product	-	Processed transcript	-	-
POLR2E-003	OTTHUMT00000458045	2441	No protein product	-	Retained intron	-	-
POLR2E-008	OTTHUMT00000458047	756	No protein product	-	Retained intron	-	-
POLR2E-005	OTTHUMT00000458050	685	No protein product	-	Retained intron	-	-
POLR2E-004	OTTHUMT00000458051	440	No protein product	-	Retained intron	-	-

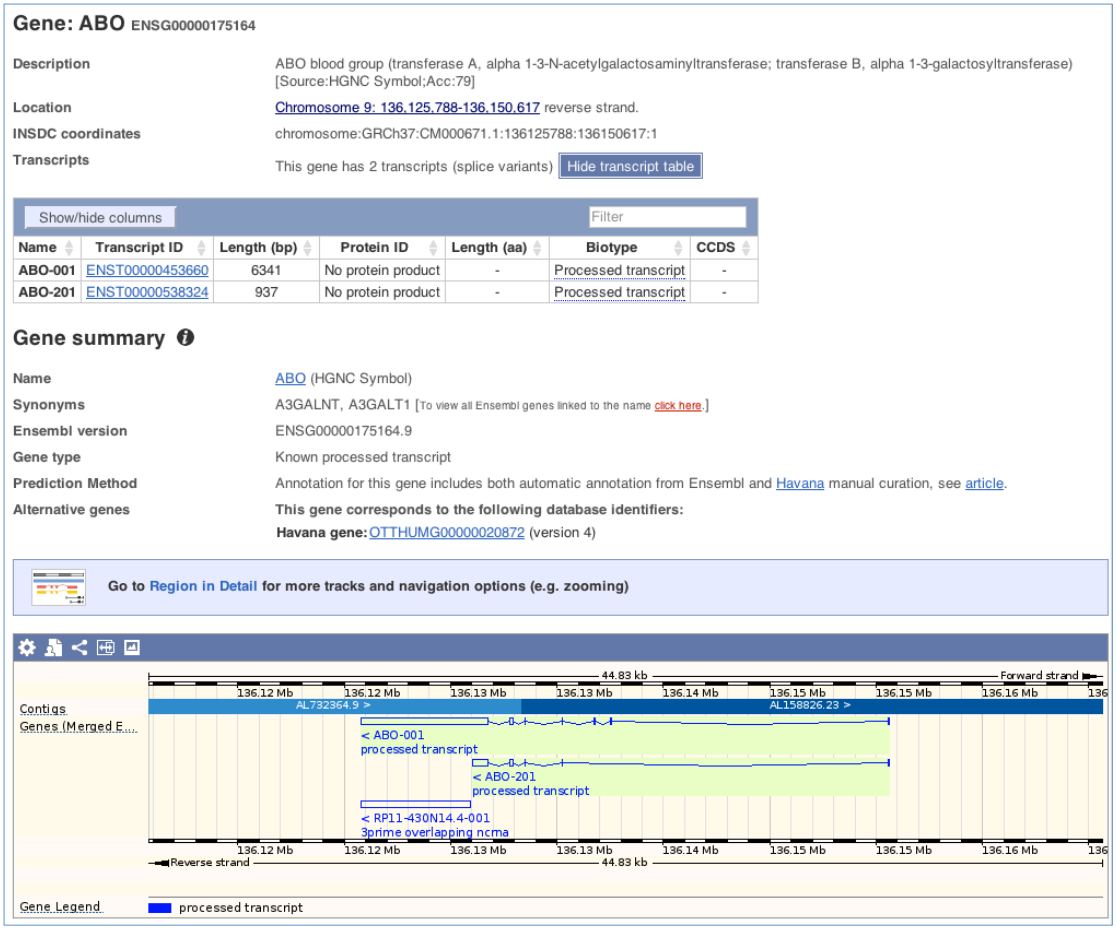
Variants 1-6 are unchanged, variant 7 is now protein coding, variant 8 is unchanged, variant 9 is now protein coding, variant 10 is unchanged. There has also been the addition of 4 new splice variants. Variant 11 is a non-coding transcript and variants 12 – 14 are protein coding.
Due to the complexity of the release process it can take up to 3 months for new annotation to be available in Vega. To address this, the Vega update track is run every two weeks for human and mouse, and so new annotation is publicly available much more quickly. These will be incorporated into the main Vega site when a there is a new human gene release (approximately every 3 months), and then later into the Ensembl merge and the Gencode geneset.

Demo: Looking at GRC patches

We're now going to look at *ABO*, a protein-coding gene known to be involved in blood grouping. From the Ensembl homepage, search for ABO in Human.



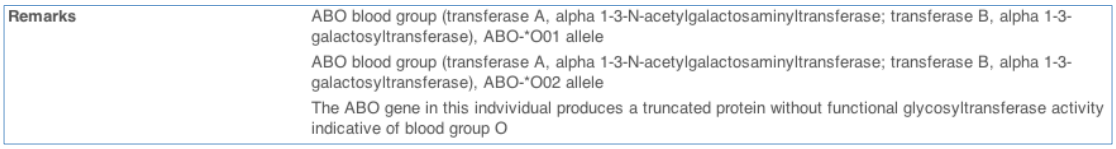
This search yields two results, the gene on the primary assembly (above) and the gene on the patch (below). To find out more about this, click on the top gene ABO (Human Gene).



This gene has two transcripts, both of which are non-coding. This does not fit with what we know about the gene, that it is protein coding.

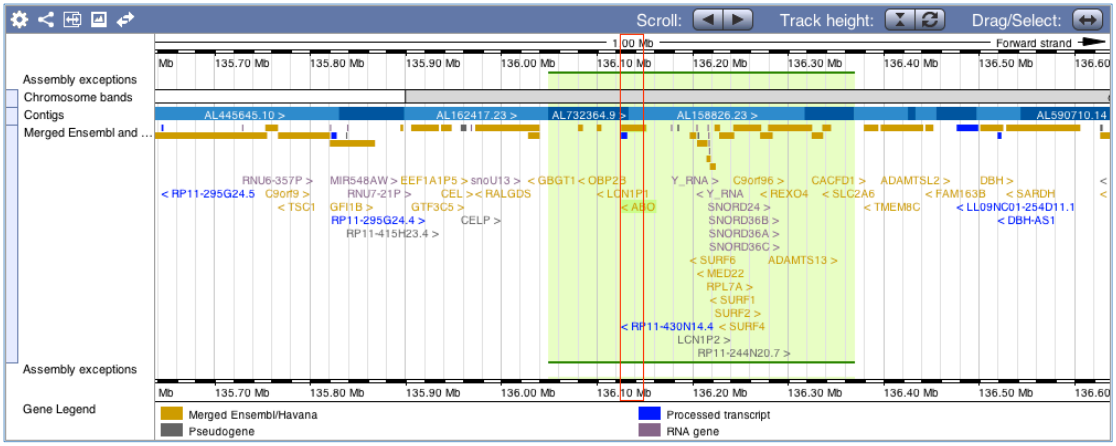
To understand what's going on, click on the Havana ID OTTHUMG00000020872 to see the gene in Vega. Open the transcript table and click on the transcript OTTHUMT00000054907.

Have a look at the Remarks:



The gene lies between 2 BAC clones and each half of the gene represents a different allele. As a result there is no coding gene for this locus.

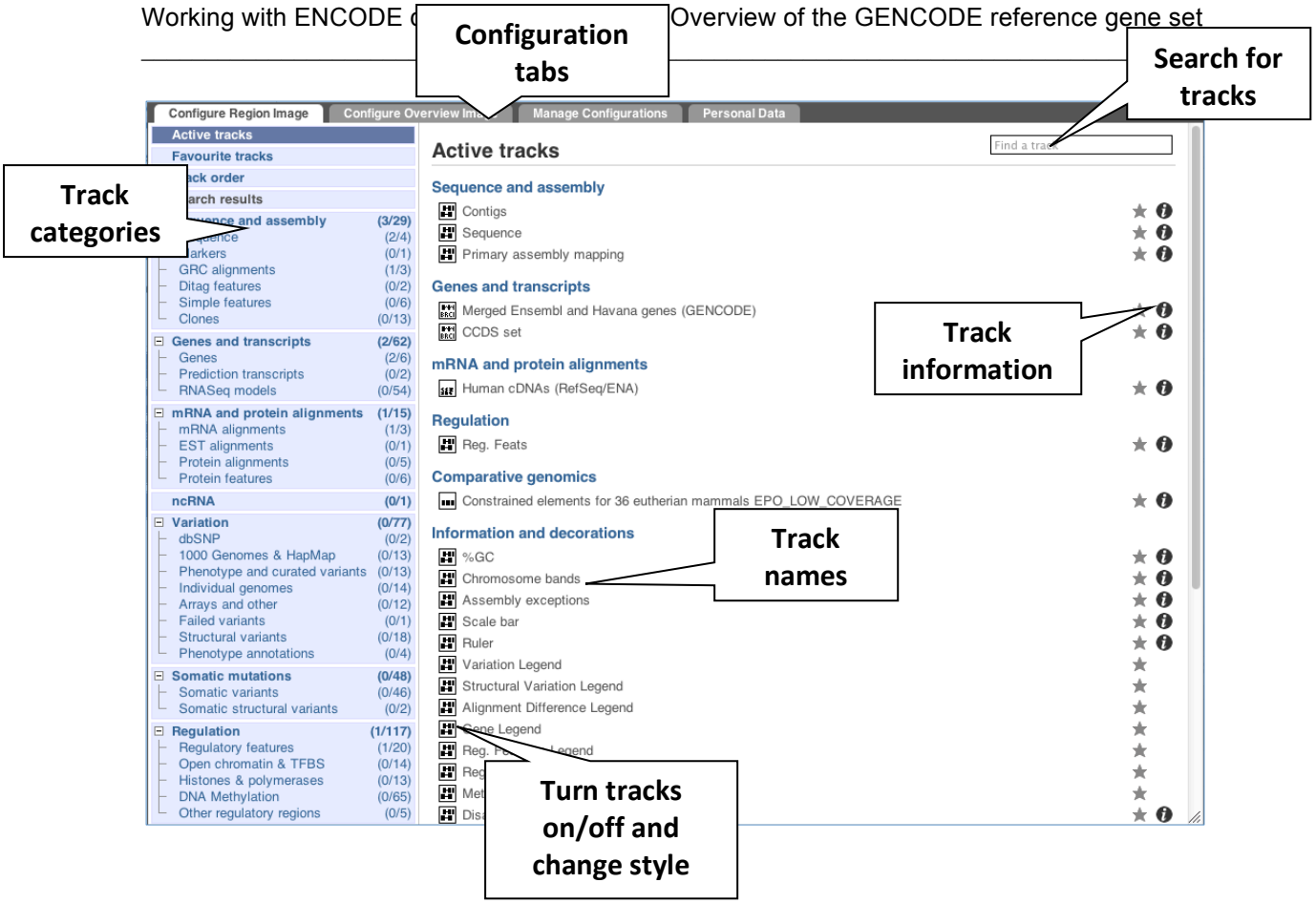
Go back into Ensembl and click onto the location tab.



As we saw in the search results, this gene falls within a patch, shown in green. To find out why, we want to add a track. By default only a very limited number of tracks is shown (note that it says at the bottom the display that 'There are currently 441 tracks turned off'). Additional tracks can be added using Configure this page.

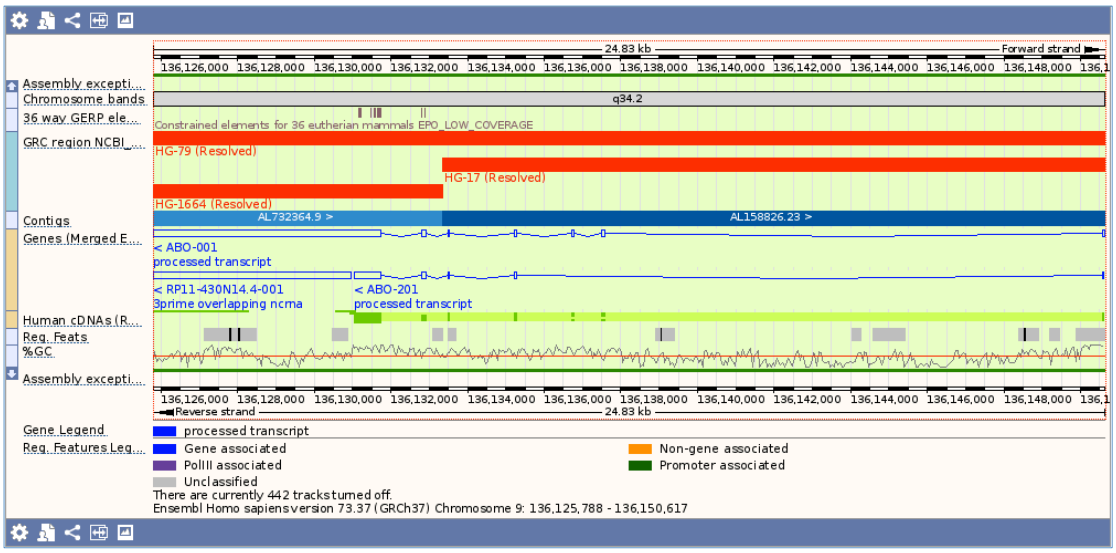
This will open a menu that allows you to change the image.

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

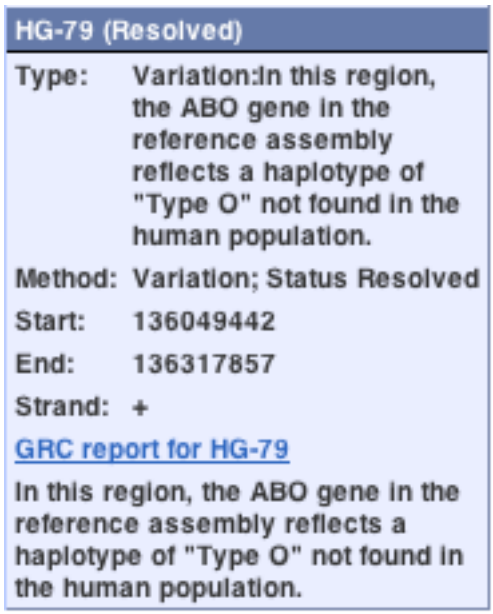


Open the Configure this page menu and select Sequence and assembly from the left. Turn on the track GRC region NCBI_37 in Labels, then close the menu by clicking anywhere outside the menu.

You can now see the track we added in red. These indicate problems in the primary assembly. All of the bars are labelled (Resolved) indicating that the problems have been fixed.

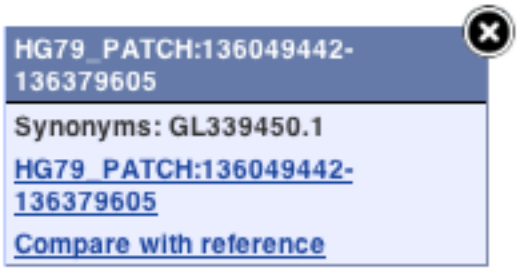


There are three red bars, one of them is labelled HG-79 (Resolved). Click on this red bar to open a pop-up.

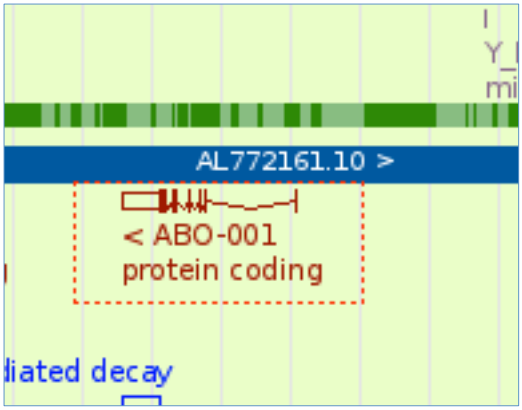


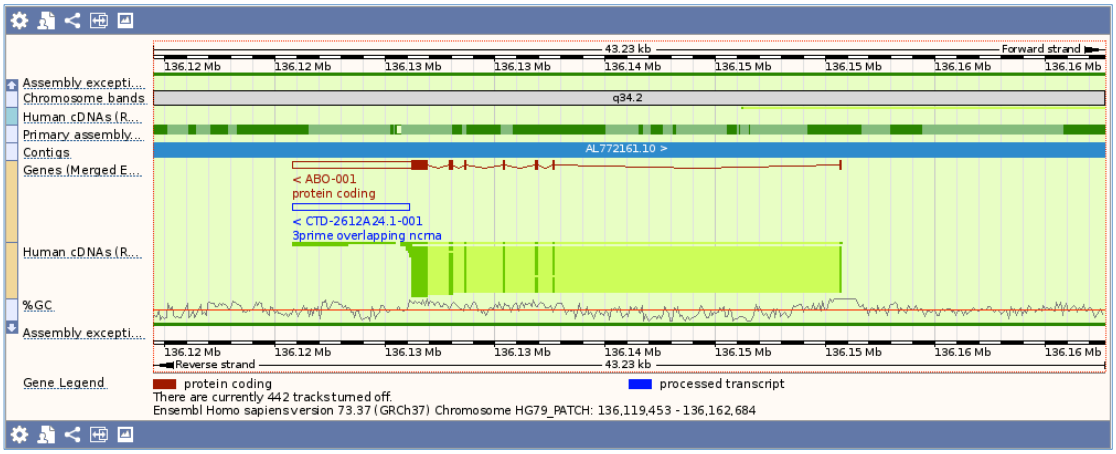
The pop-up tells us that the reference assembly gives us a gene that doesn't exist in human populations. This is due to the fusion of two alleles. Since this is listed as Resolved, we know that it was fixed by a patch.

The thin dark green line at the top of the image indicates the position of the patch. Click on it to open a pop-up menu.



Click on HG79_PATCH:136049442-13637605 to go to the Region in Detail view for the patch. We have zoomed right out to view the whole patch. We can zoom back in on the ABO gene by dragging out a box around it.





We can now see the ABO gene labelled as protein_coding. We can also see that the whole of the gene is covered by a single clone. Using a single clone meant that all of the data came from a single genomic sequence, so the gene fusion problem is resolved.

Click on the link to UCSC at the left to open the patch in UCSC.

It is not possible to view patches in UCSC because they only have the primary assembly.

Exercises

Exercise 1 – Searching for splice variants of a gene

Search for the *BRAF* gene in human in Ensembl, which is an important gene in cancer. How many splice variants are there and what are their biotypes? How many merged transcripts are there? How many have a CCDS?

Exercise 2 – Searching for genes on haplotypes

(a) Search for the *HERC2* gene in human in Ensembl. How many genes are there called *HERC2*? Why is this? Take a look at both genes. Is there a difference between the number of transcripts?

(b) What is the name of the haplotype that the alternate sequence falls on? Go to the location view. Can you compare the haplotype with the primary assembly? What differences can you see?

Answers

Exercise 1 – Searching for splice variants of a gene

Go to ensembl.org.
Select human from the drop down list and type in braf then hit return.
Click BRAF (Human Gene).
If the transcript table is hidden, click on Show transcript table.

There are five transcripts of *BRAF*. Two are protein coding, two are subject to nonsense-mediated decay and one has a retained intron. One has a CCDS.

Exercise 2 – Searching for genes on haplotypes

(a) Go to ensembl.org.
Select human from the drop down list and type in herc2 then hit return.
Narrow down to genes only by selecting Gene from the left hand list.

The search returns 11 genes in total, of which two are called *HERC2*. One of these is on a haplotype, where the genome has different sets of variants between individuals, and the other is on the primary sequence.

Open the two genes, HERC2 (Human Gene) and HERC2 (Human Alternate sequence Gene), in different tabs.

The primary assembly gene has twelve transcripts, whilst the haplotype gene has eight.

(b) The alternate sequence gene is described as being on HRSCH15_1_CTG4.

Click on the Location tab in the top bar from either gene. You will see the haplotype represented as a red highlighted region. Click on the dark red line at the top or bottom of the region, then select Compare with patch or Compare with reference, depending on which one you're looking at.

There is a short intergenic insertion in the haplotype compared to the primary assembly.