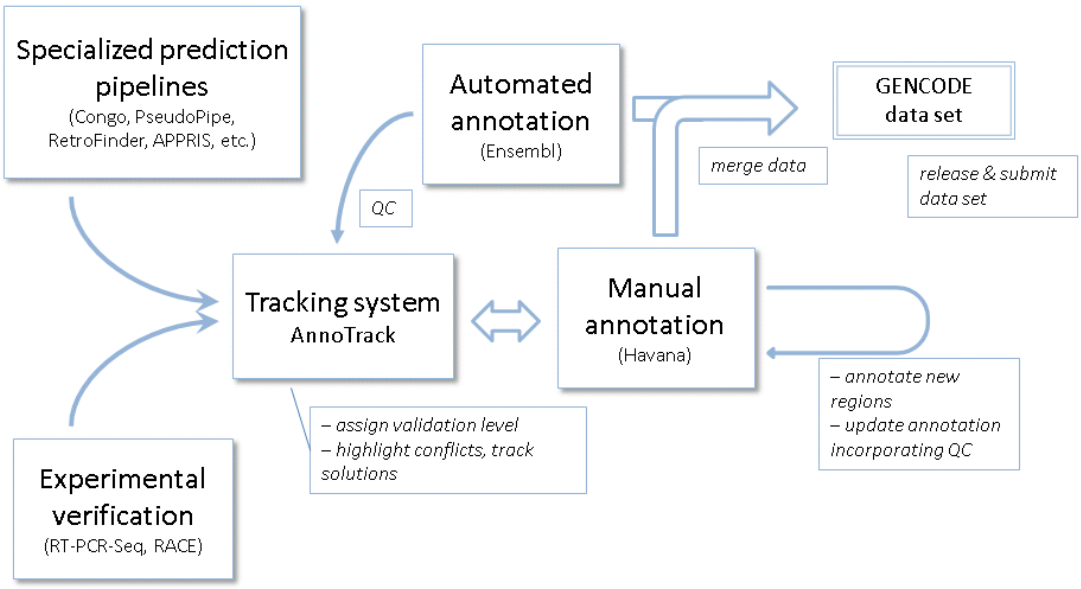
Overview of the GENCODE reference gene set

Aims

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

Introduction

Schematic showing interconnection between different GENCODE pipelines

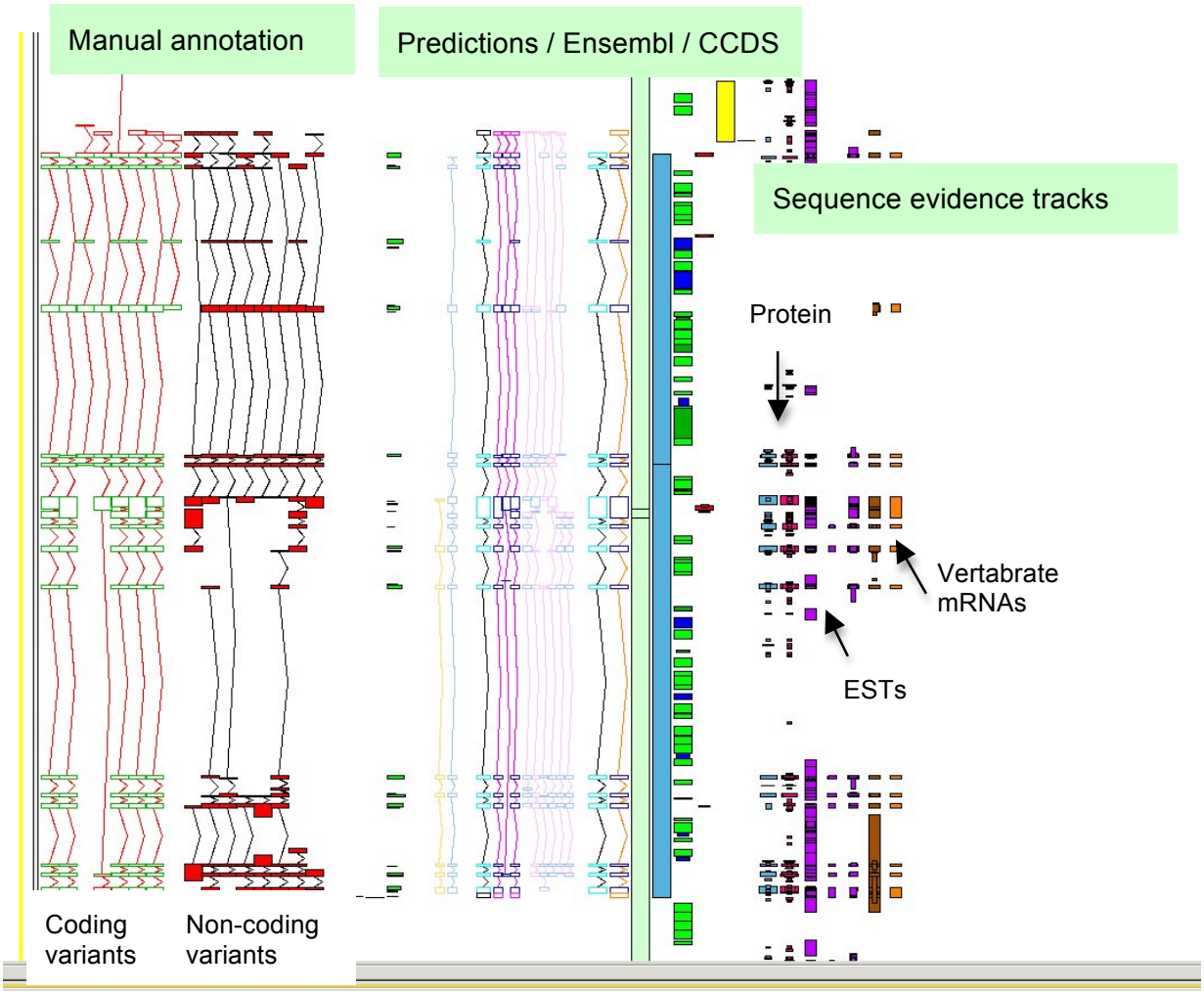


HAVANA (Human and Vertebrate Analysis and Annotation) group at the WTSI perform manual genome annotation. Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases (including cross-species) and a series of *ab initio* 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).

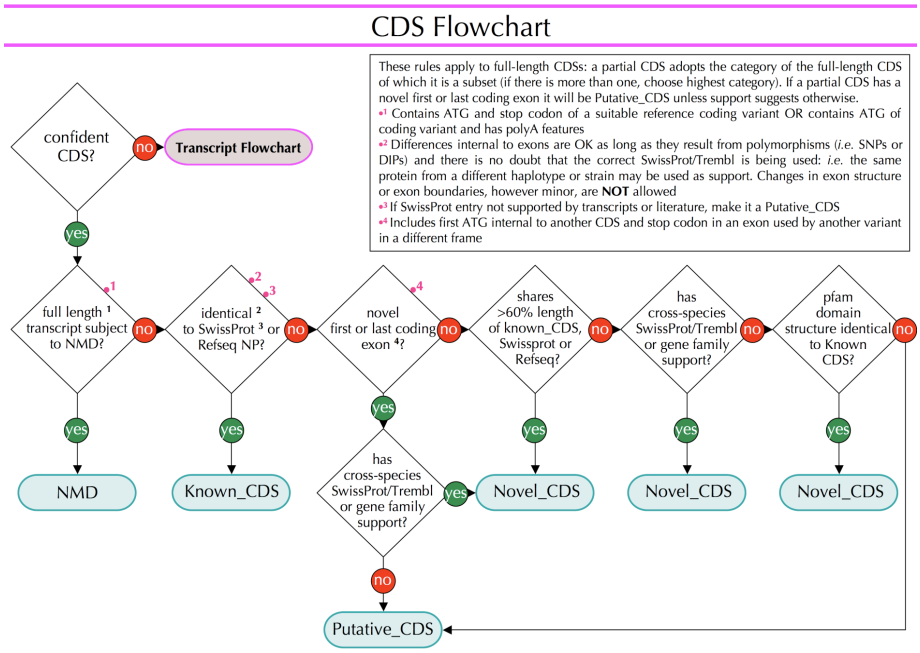
Manual Genome Annotation

Genomic sequence is run through the analysis pipeline and saved in the mysql database. The annotators then view this data through the Zmap viewer and perform manual annotation in the Otterlace transcript editing interface. The annotation is then save back to the database. Every few months this data is fed through to Vega and then also incorporated into the Ensembl genebuild. The underlying data for the Vega database is generated by the Havana group. Vega may be browsed and searched in a similar way to Ensembl.

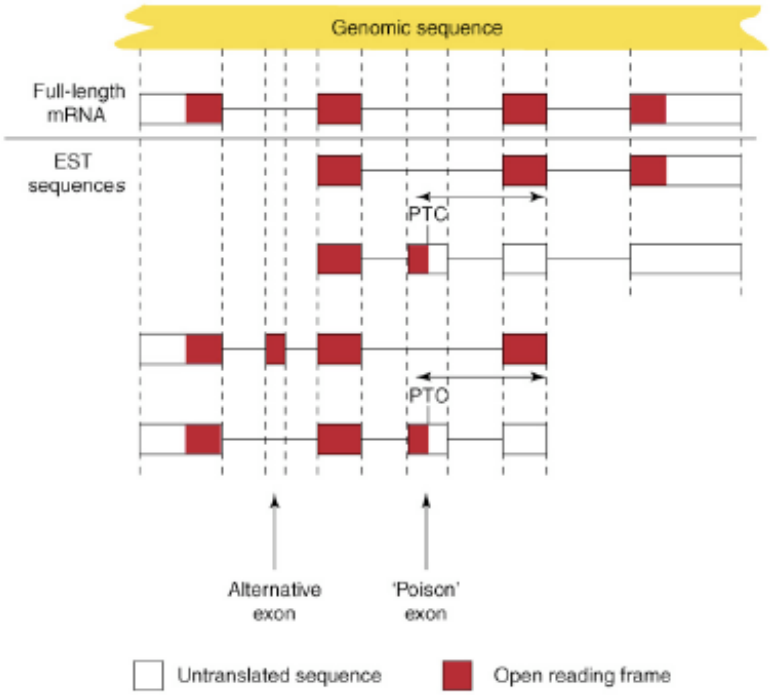
Below is a screen-shot of the CIZ1 locus in Zmap from the Otterlace annotation software. Protein coding genes are shown in red and green, whilst non-coding transcripts are shown in red.
Other columns show Blast hits to DNA and protein databases, repeats and Phastcons regions (evolutionarily conserved regions from 28 vertebrates)



Biotypes: The Havana team annotate both coding and non-coding loci, including pseudogenes.



We also annotate transcripts that are likely to be subject to nonsense-mediated decay (NMD) (PMID: 19543372, 12502788) with an intact CDS.



The exact mechanisms behind NMD have not been elucidated and so we retain the CDS in our gene models.

The Vega database (http://vega.sanger.ac.uk/)

The Vertebrate Genome Annotation (Vega) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence. Vega differs from Ensembl in that it shows annotation from the labour intensive process of manual curation produced by the HAVANA (Human and Vertebrate Analysis and Annotation) group at the WTSI. Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases and a series of *ab initio* gene predictions. Annotation is based on supporting evidence, which is external sequence such as ESTs, cDNAs and protein, and is performed to standards guidelines available from described in the HAVANA annotation manual (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/).

Vega displays complete chromosome regions in blue and dark grey showing regions with no annotation.

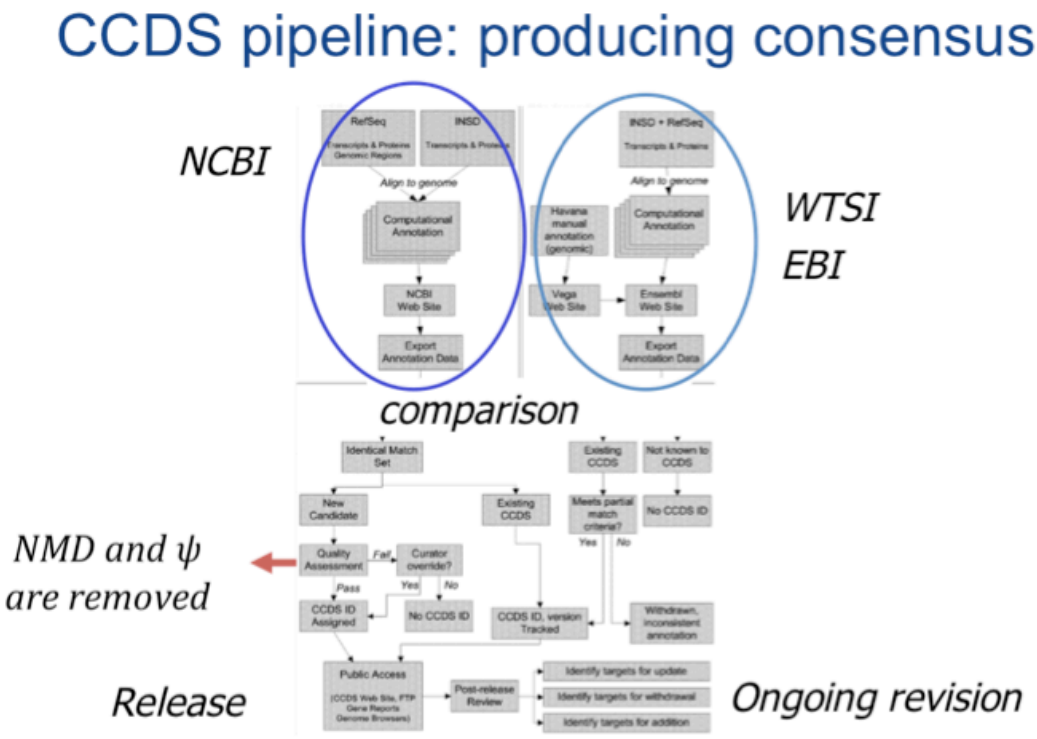
Major Histocompatability Complex

The human major histocompatibility complex (MHC) contains many immune related genes including highly polymorphic examples encoding MHC class I and class II molecules that present antigens to T lymphocytes. Vega has seven human haplotypes of the chromosome 6 MHC region together with reference sequence 6-PGF: 6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF. These are shown as distinct chromosomal regions and are also included in the Vega comparative analysis.

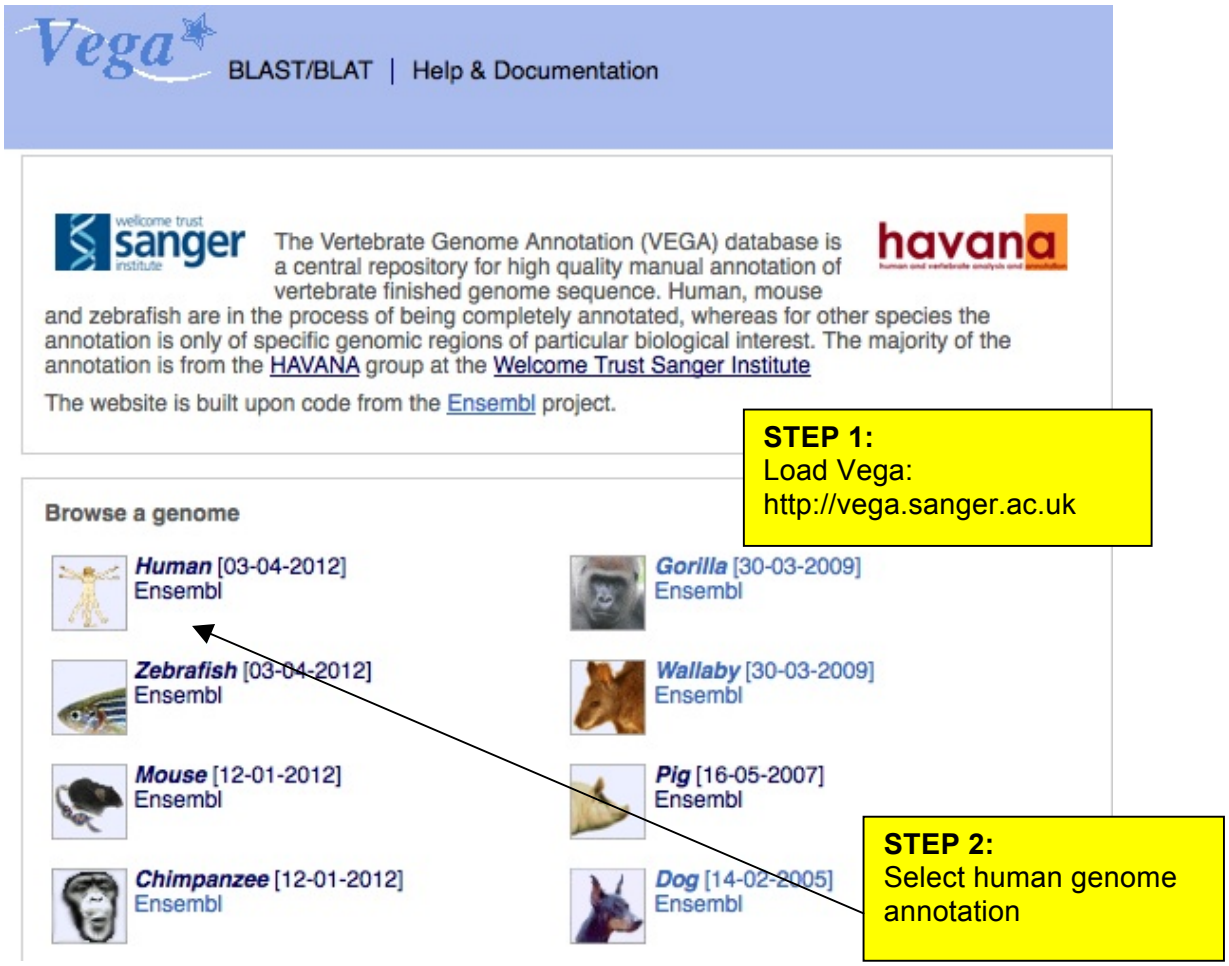
CCDS

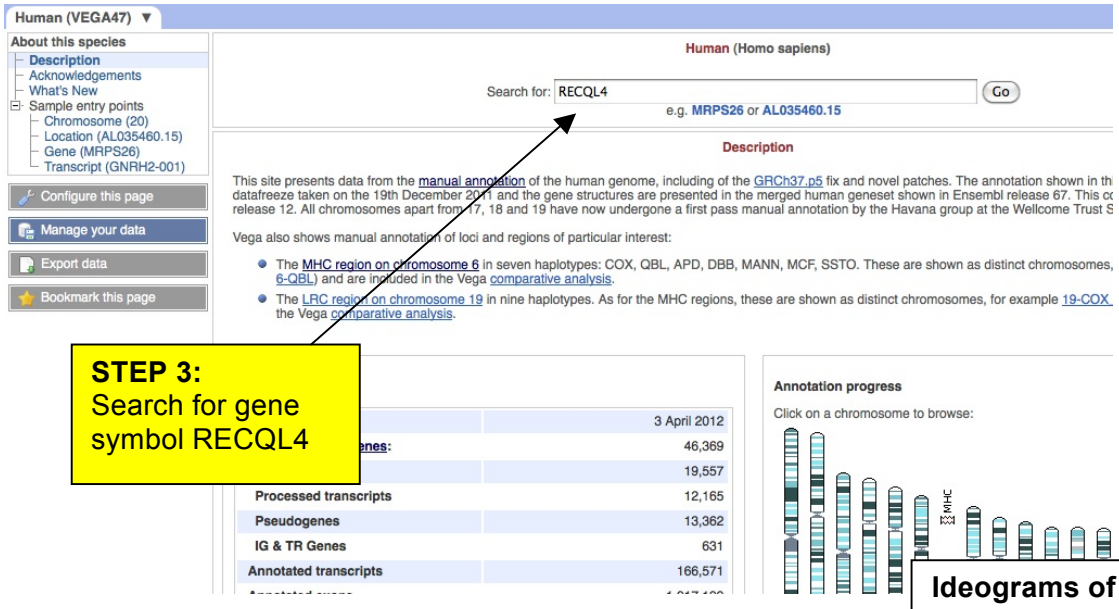
HAVANA is an important contributor to the Consensus CDS (CCDS) project, which is a collaborative effort between the European Bioinformatics Institute (EBI), the National Centre for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI) and the University of California at Santa Cruz (UCSC). The aim of the project is to identify a core set of human protein coding regions that are consistently annotated between the different

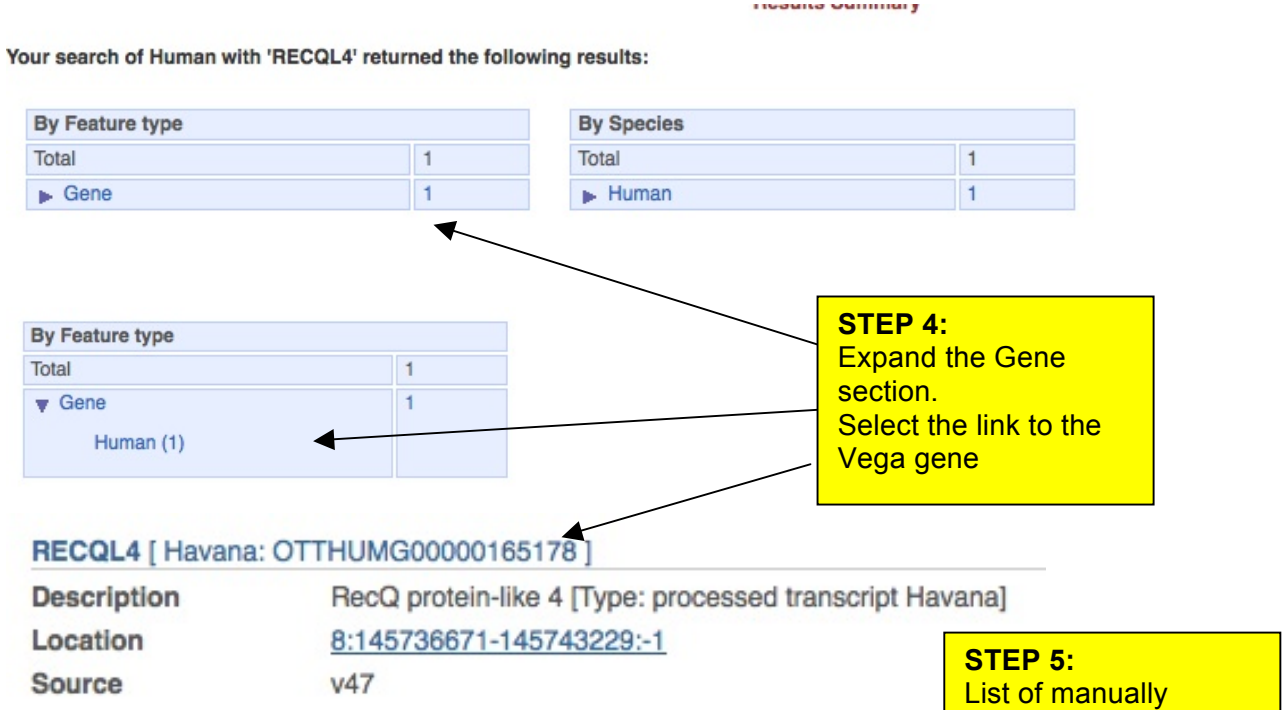
institutes. The long-term goal is to support convergence towards a standard set of gene annotations. The CCDS gene set is generated by Ensembl and NCBI and there is extensive QC by WTSI, NCBI and UCSC. A set of guidelines have been developed for the annotation of coding sequence regions by the collaborating Institutes, and any changes to the CCDS set have to be agreed by all three sites.



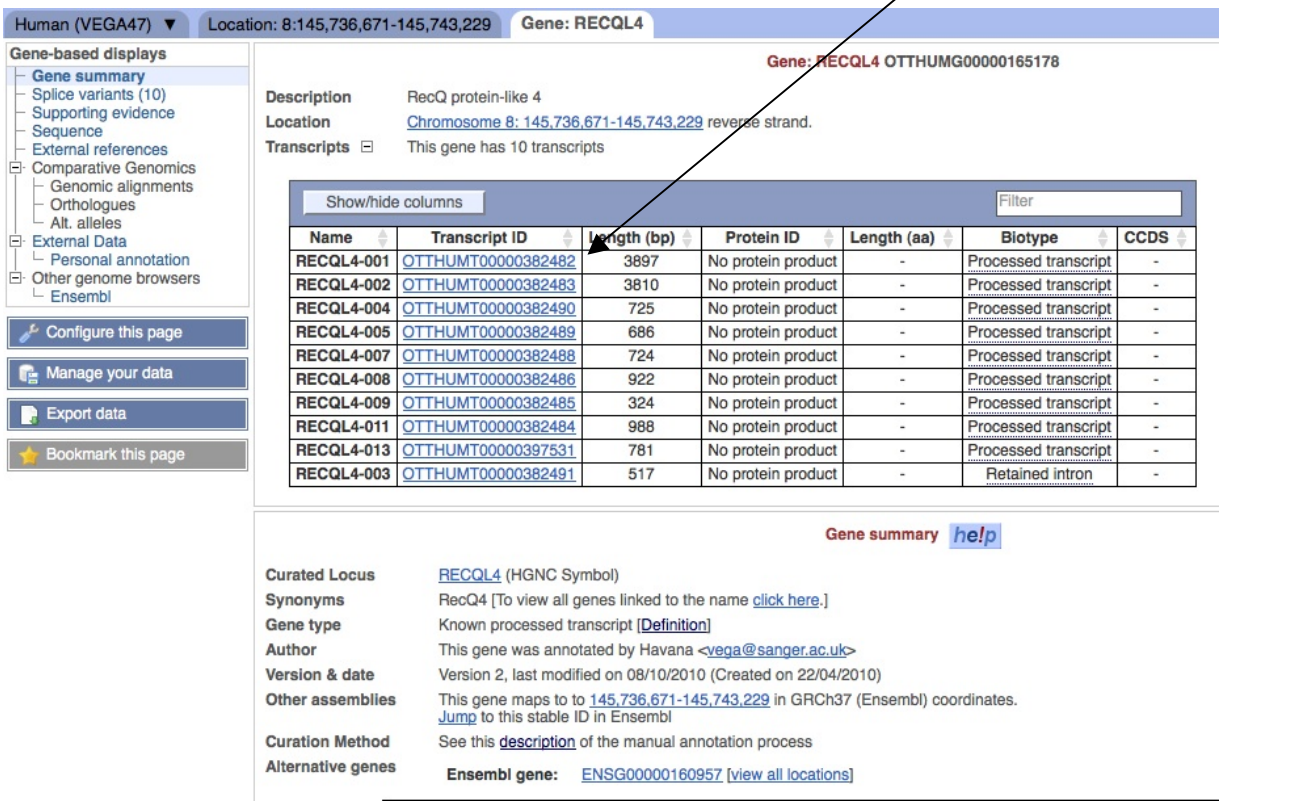
Worked example 1:
View the RECQL4 locus. What biotype is this gene in Vega, Ensembl and UCSC?

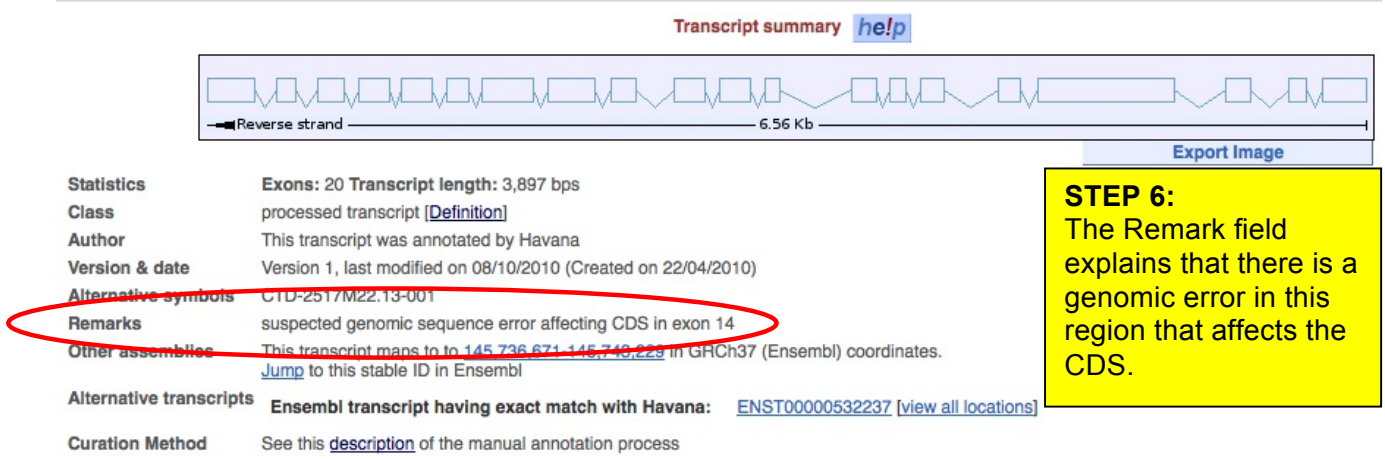






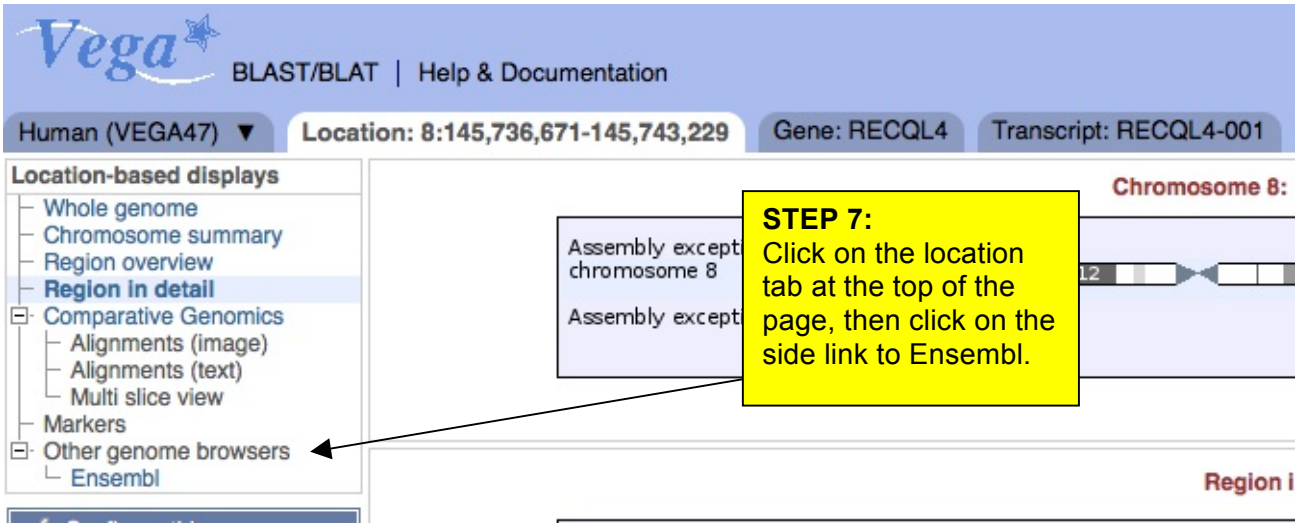
Notice that there are no protein products for this gene.





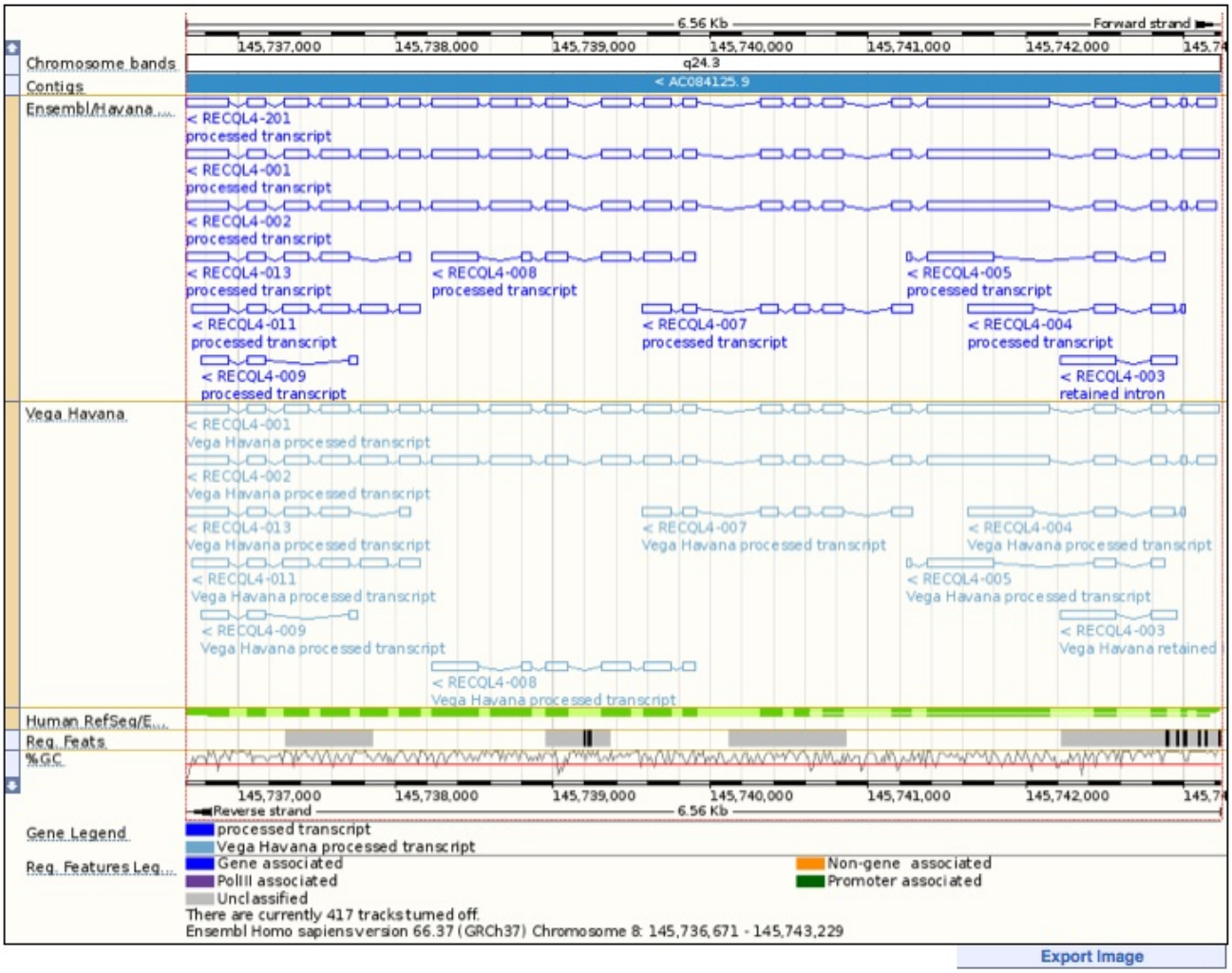
STEP 6:
The Remark field explains that there is a genomic error in this region that affects the CDS.

As there is a suspected genomic error we should check and see if this is being investigated by the GRC. In order to view the GRC track in a genome browser we will need to go to Ensembl.



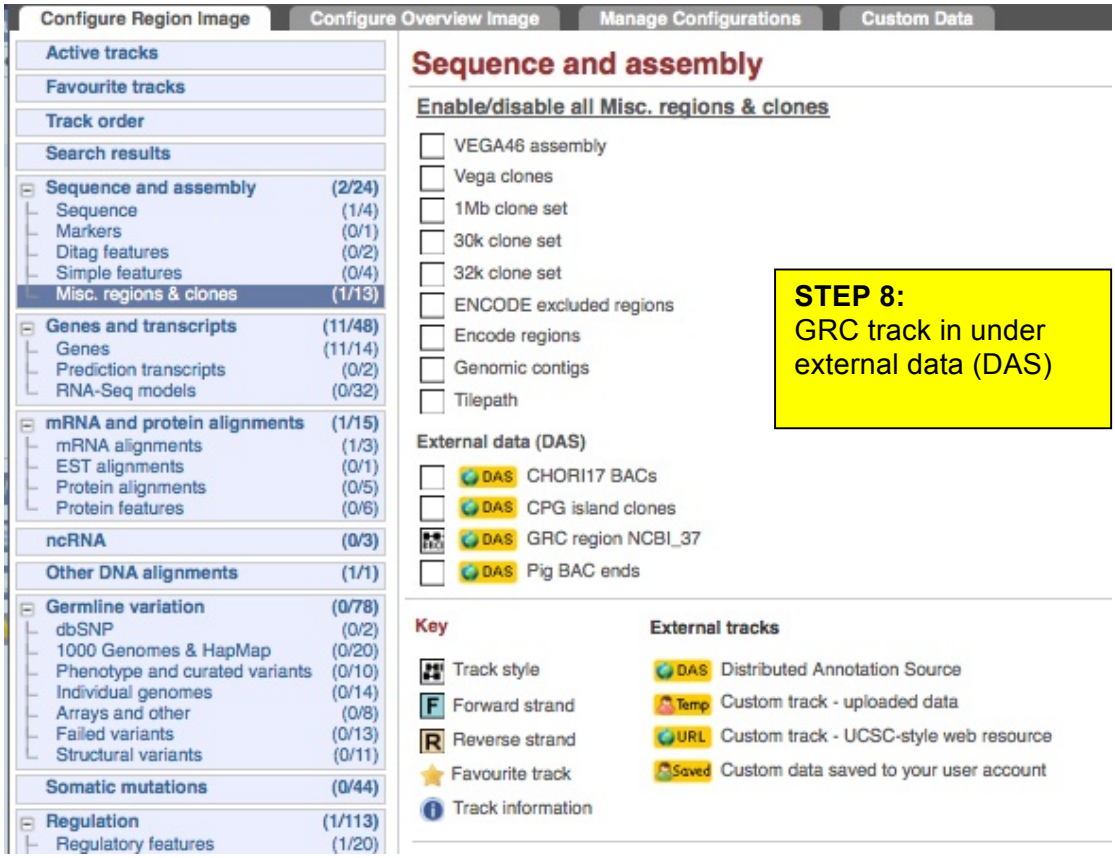
STEP 7:
Click on the location tab at the top of the page, then click on the side link to Ensembl.

This will bring you to the same gene in the Ensembl genome browser, which is also displayed as non-coding.

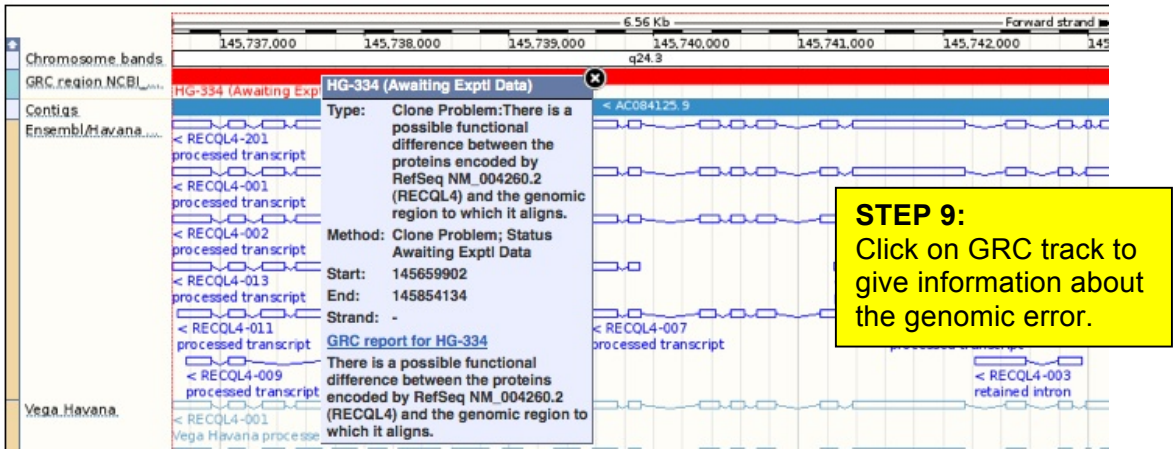


This is the default view, but there are many tracks that you can switch on and will be expanded on in a later module. In order to view the GRC track you will need to go to

Under “Sequence and assembly” select “Misc. regions and clones”. Then under “External data (DAS)” you will see “GRC region NCBI_37”. Select this with labels:

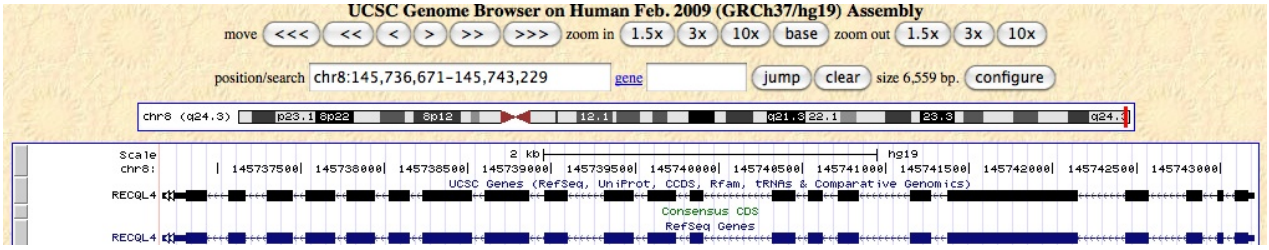


The GRC track is shown in red, if there is a GRC report for that region.

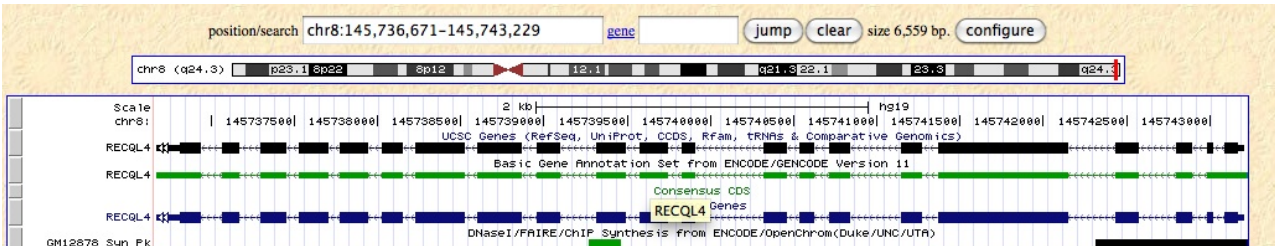


We can now jump to the same gene in the UCSC genome browser, by clicking on the UCSC link in the side bar

This will open in a new browser window.

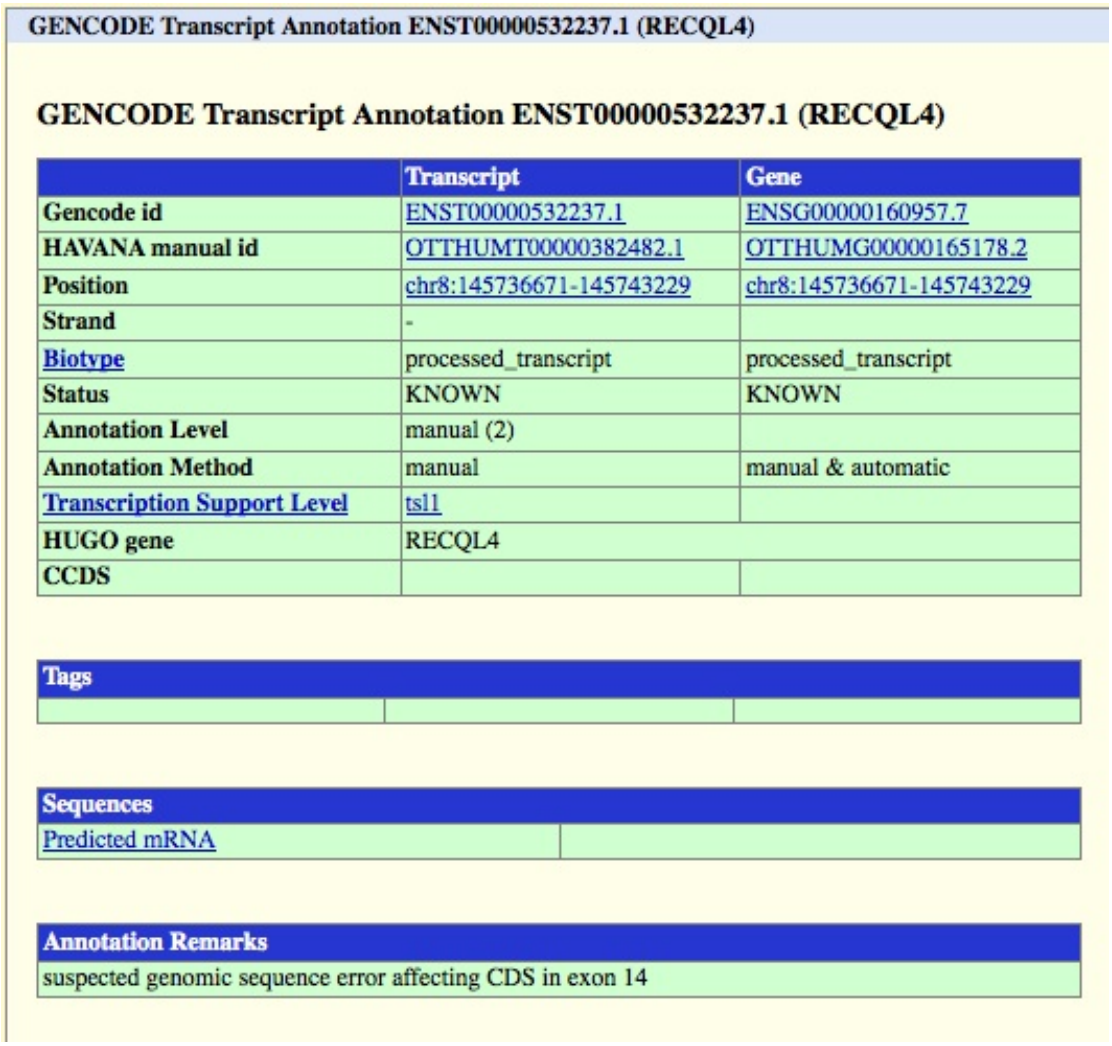


Switch on the GENCODE genes V11 and mouse over the track:



STEP 10:
Mouse over the GENCODE V11 track and when the gene name pops up, click on it to open a new window that explains the track.

The annotation remarks from the manual annotation can be viewed with lots of other information describing the track.



GENCODE Transcript Annotation ENST00000532237.1 (RECQL4)

GENCODE Transcript Annotation ENST00000532237.1 (RECQL4)

	Transcript	Gene
Gencode id	ENST00000532237.1	ENSG00000160957.7
HAVANA manual id	OTTHUMT00000382482.1	OTTHUMG00000165178.2
Position	chr8:145736671-145743229	chr8:145736671-145743229
Strand	-	
Biotype	processed_transcript	processed_transcript
Status	KNOWN	KNOWN
Annotation Level	manual (2)	
Annotation Method	manual	manual & automatic
Transcription Support Level	tsl1	
HUGO gene	RECQL4	
CCDS		

Tags		

Sequences	
Predicted mRNA	

Annotation Remarks
suspected genomic sequence error affecting CDS in exon 14

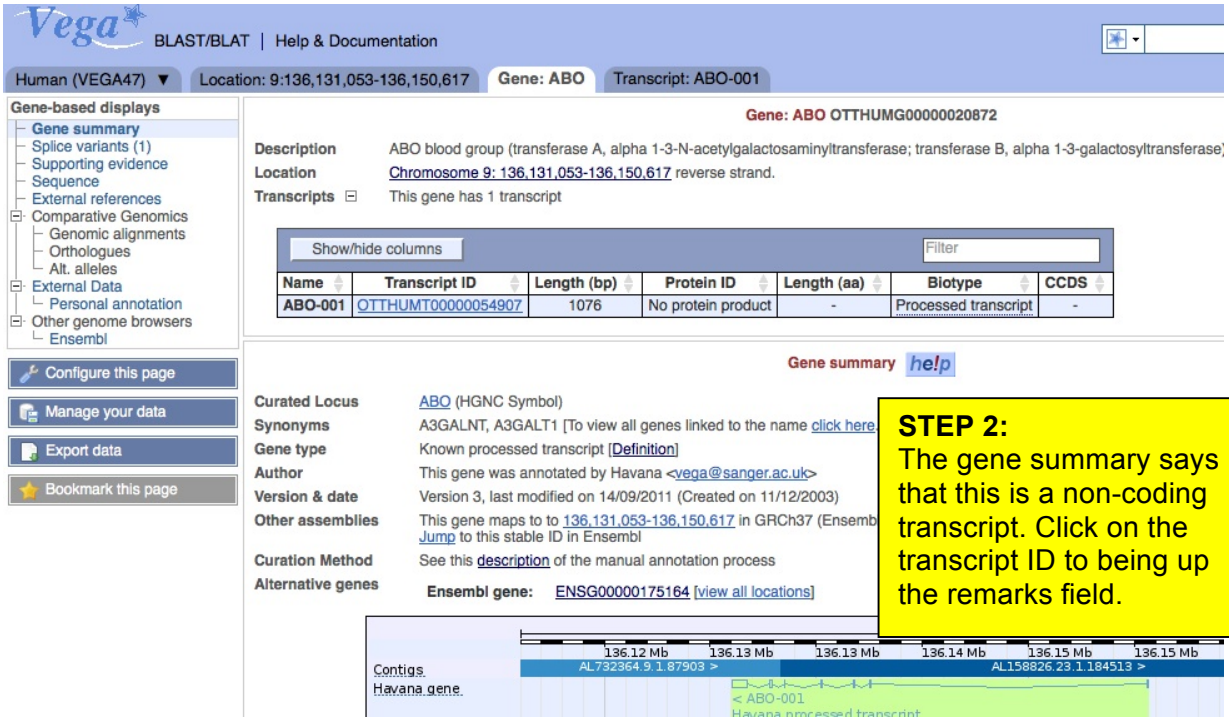
The screen shot has been truncated to save space.

Worked Example 2:

Viewing GRC patches. Look at the ABO gene.

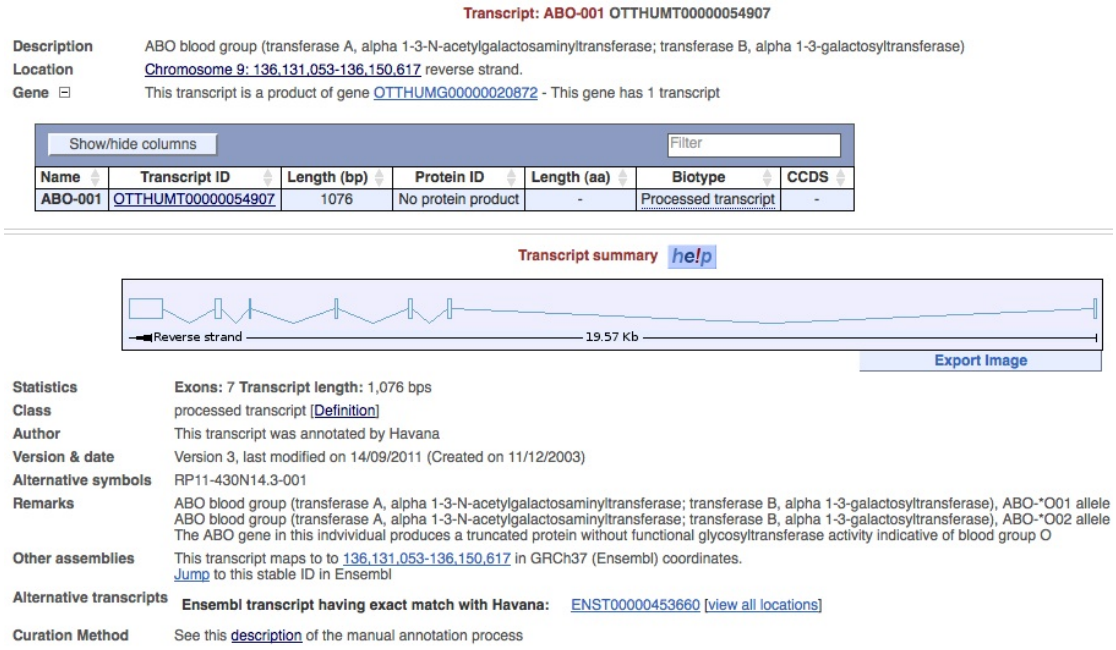
STEP 1:
Search for the ABO gene in human Vega. There are 2 hits for the same gene symbol, but they have different locations as one of them is on a GRC patch.

View the reference genome hit for ABO and look at the biotype of the gene.



STEP 2:
The gene summary says that this is a non-coding transcript. Click on the transcript ID to being up the remarks field.

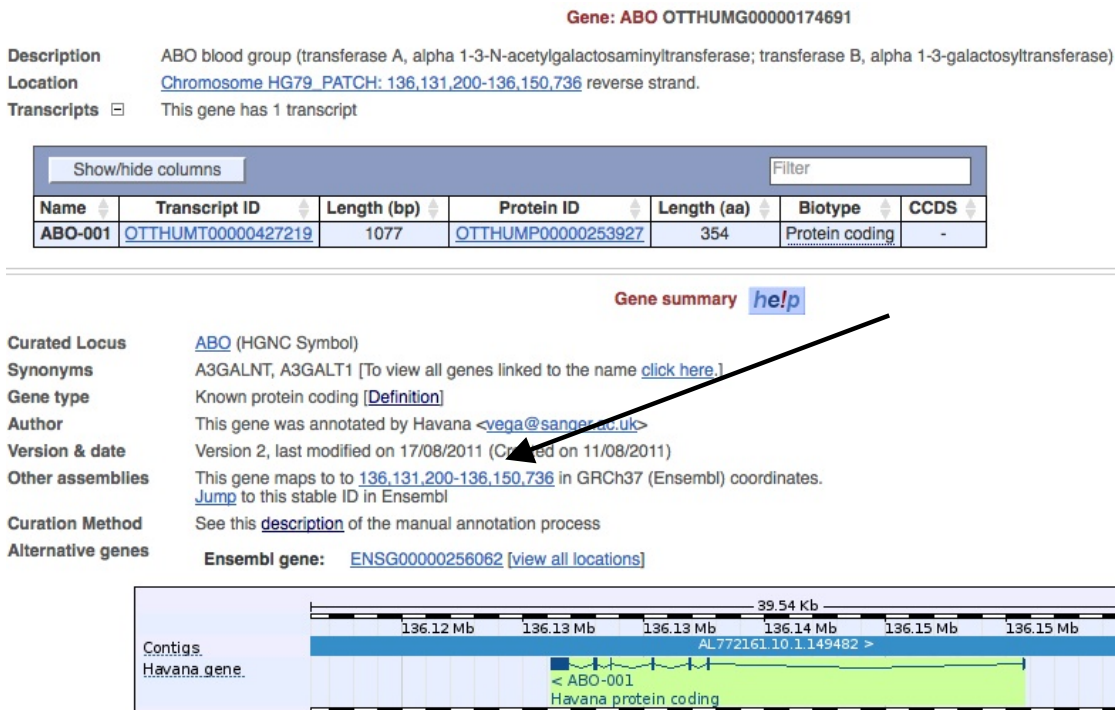
The gene is a non-coding transcript.



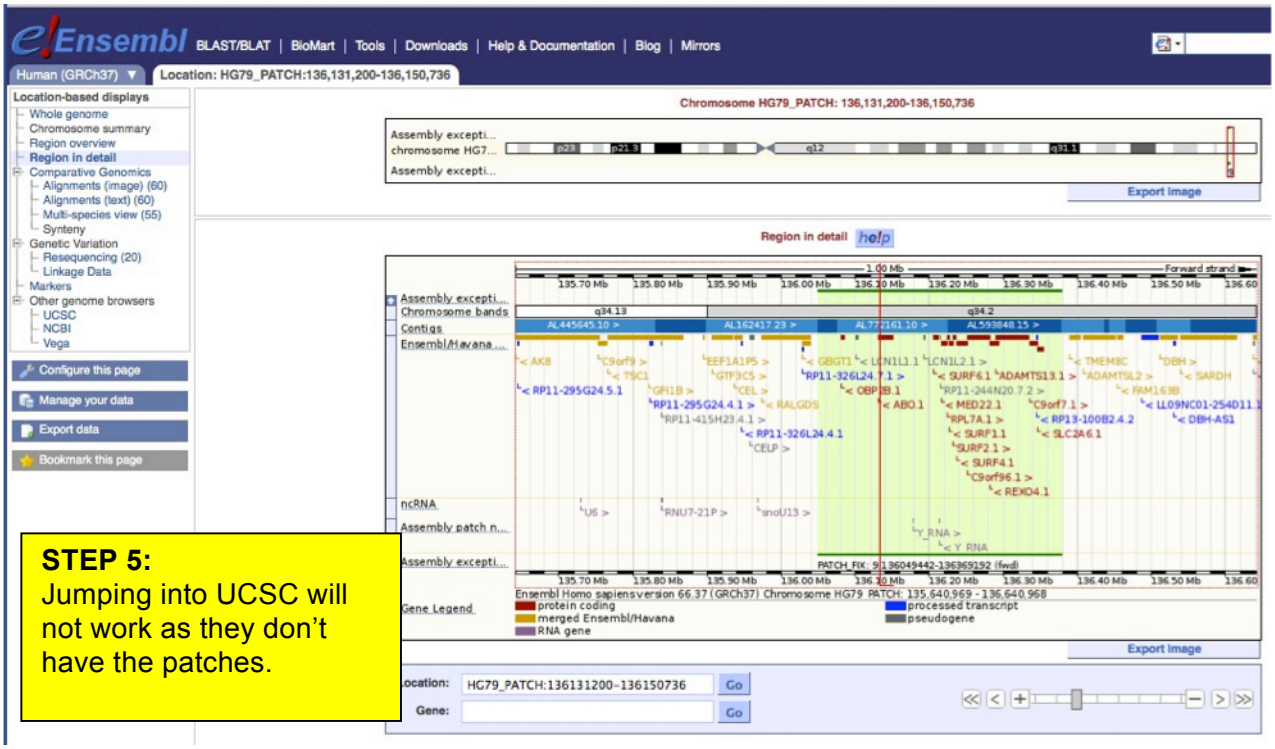
The gene lies between 2 BAC clones and each half of the gene represents a different allele. As a result there is no coding gene for this locus.

STEP 3:
Click onto the gene ID for the HG_79 PATCH entry.

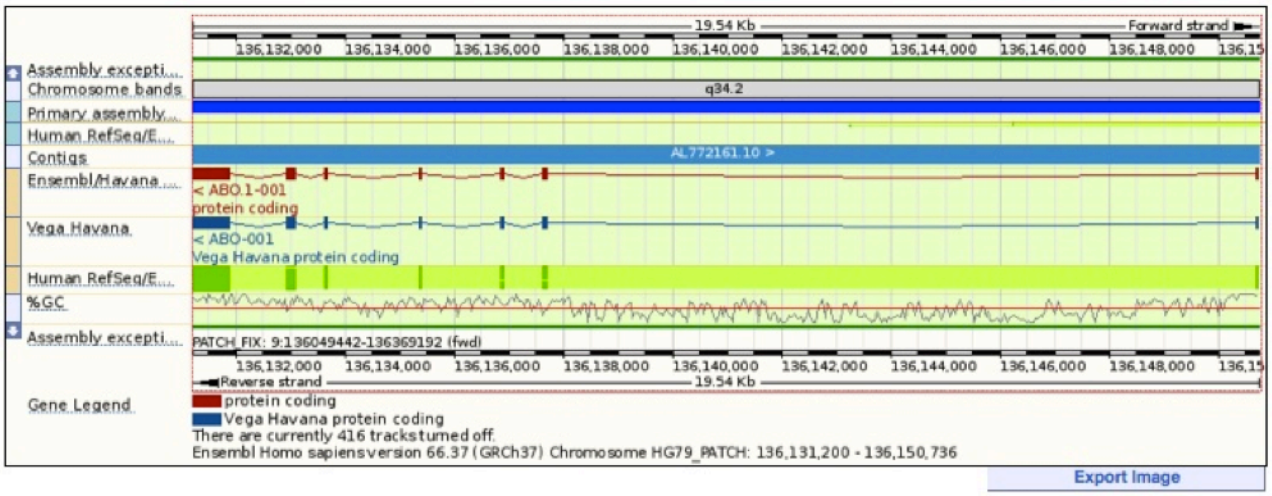
The gene is now protein coding in the patch assembly:

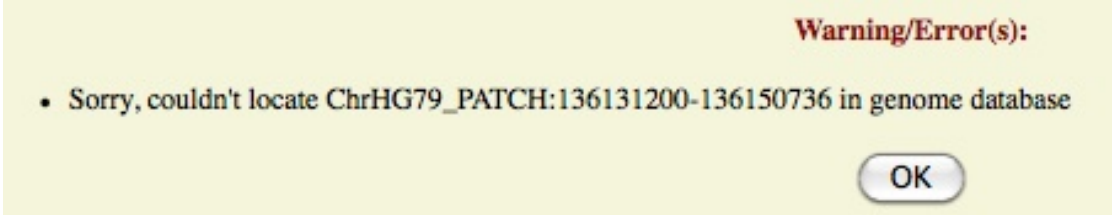


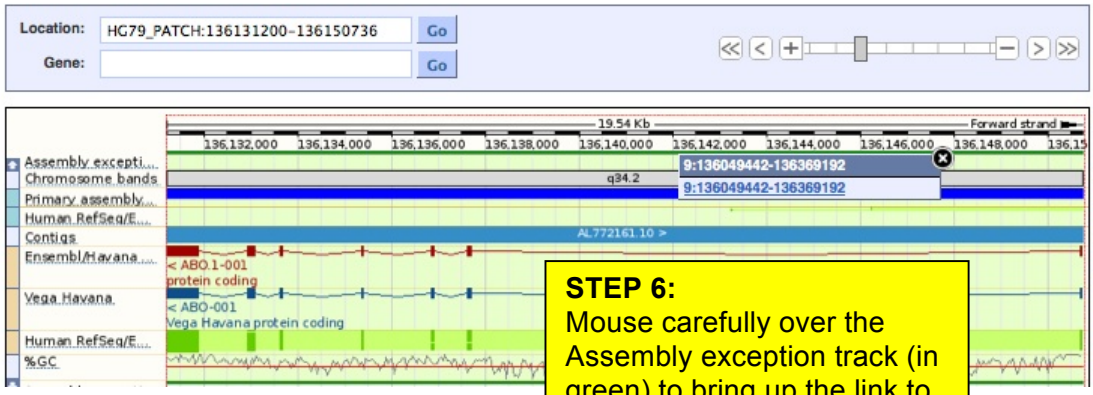
Patch assembly in Ensembl:



STEP 5:
Jumping into UCSC will not work as they don't have the patches.

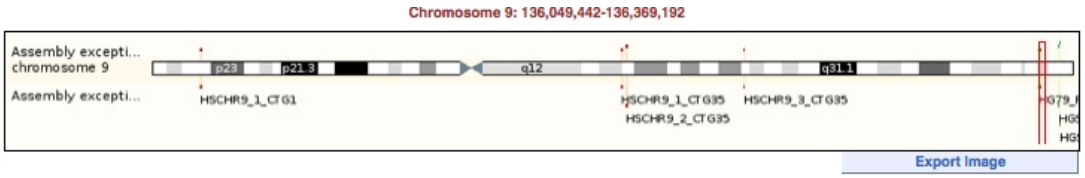


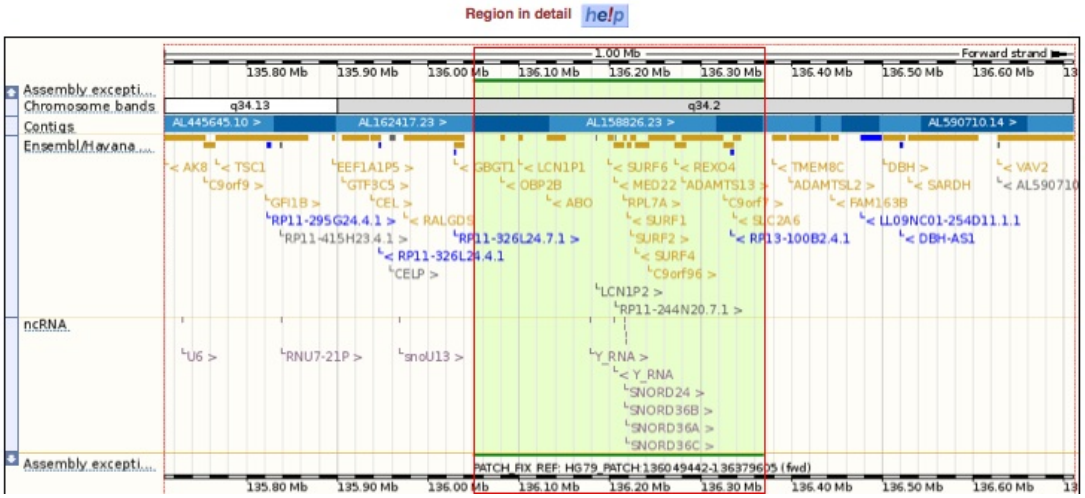




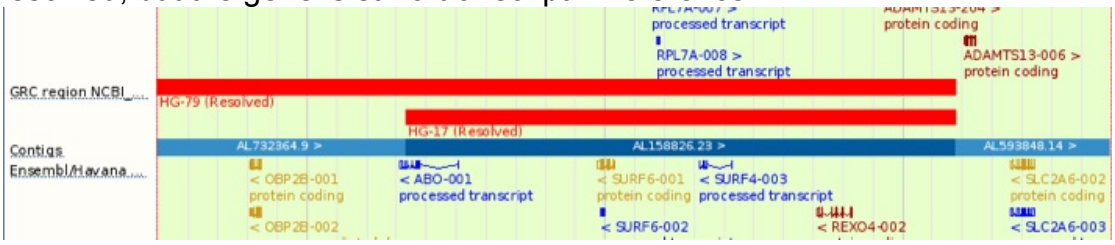
STEP 6:
Mouse carefully over the Assembly exception track (in green) to bring up the link to the reference assembly. Click on this to take you through to this gene in the reference assembly.

We are now back in reference human genome assembly.

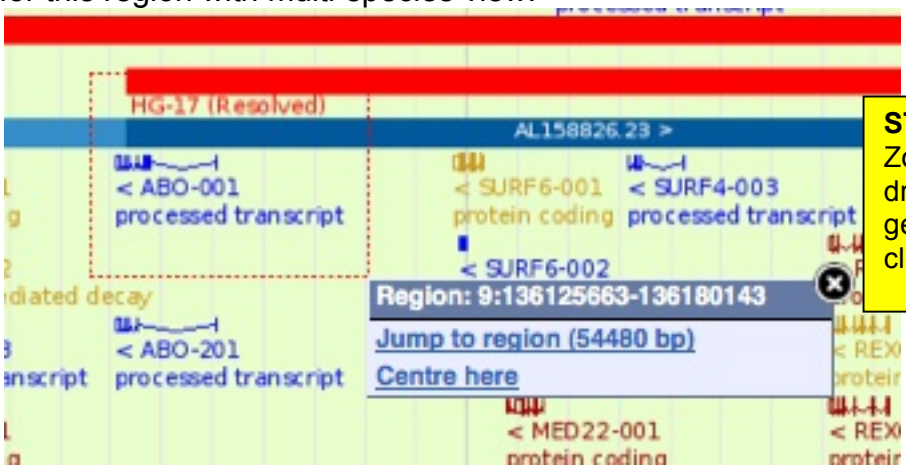




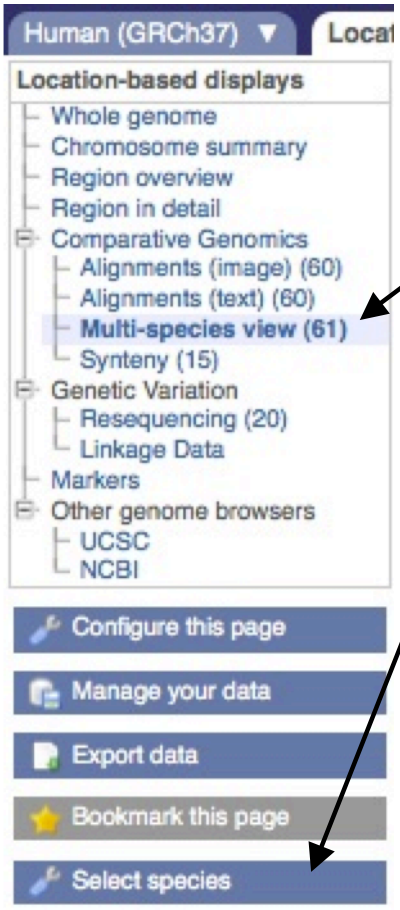
Scrolling down the page reveals that the genomic error in this region has been resolved, but the gene is still a transcript in reference.



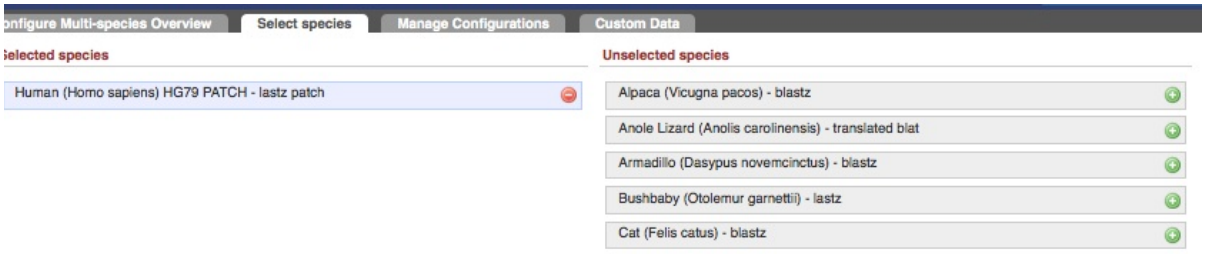
You may view the alignment between the reference and PATCH assemblies for this region with multi-species view.



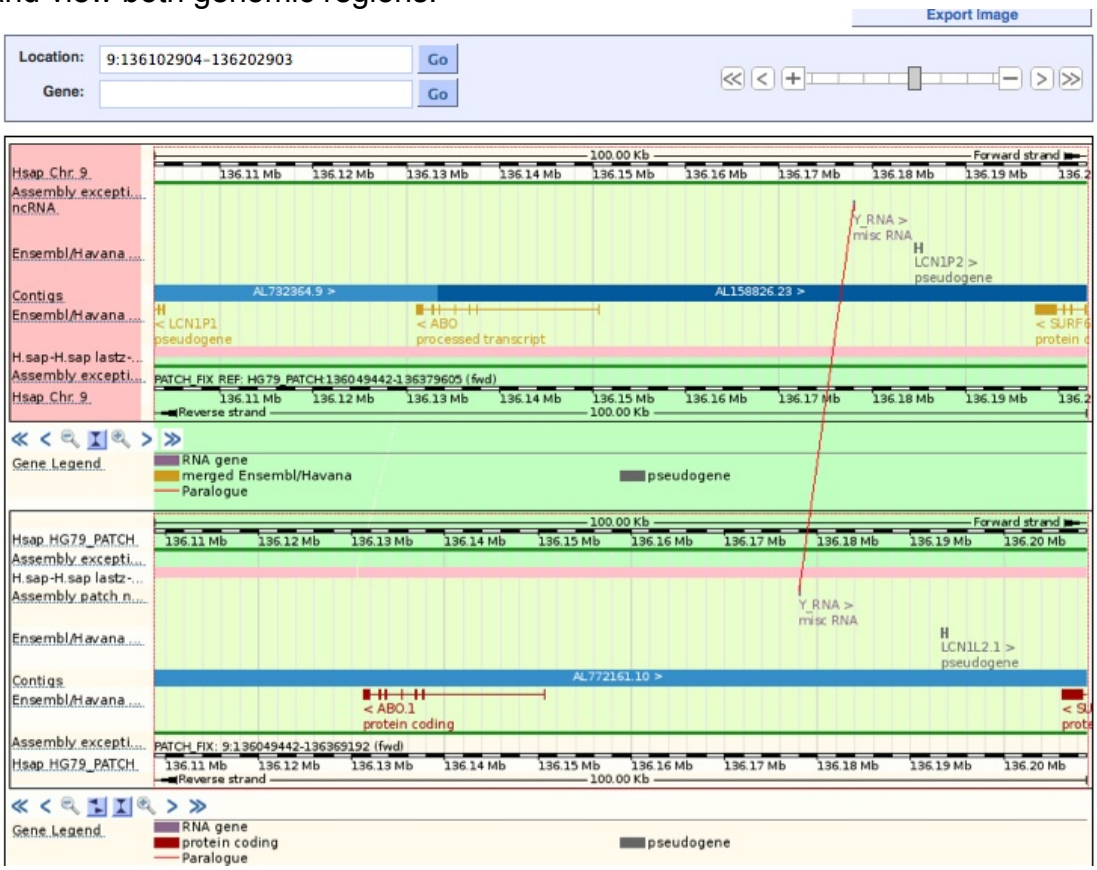
STEP 6:
Zoom into the ABO gene by drawing a box around the gene with your mouse and click on “Jump to region”.



STEP 7:
Select multi-species view. Then click on “Select species” and click on Human HG79 PATCH and it will move from the “Unselected species” column to the “Selected species” column.



Zoom out a few clicks with the slider and view both genomic regions.



The top part of the display is the reference assembly, showing the gene crossing the boundary between two BACs, and the bottom part is the PATCH alternative assembly.

Tasks

1.
Search for the FGCR2 gene in Vega.
How many alternative variants are there and what are their biotypes?

2.
Search for the HERC2 gene in Vega.
How many entries do you get from the search and why?
Take a look at the reference assembly gene. How many alternative variants are there and what biotypes are they?
Which strand is this gene located on?

3.
Zoom out a little to view the region upstream of this gene in the two neighbouring clones. Change your view to incorporate these two clones.
What is the name of these two BAC clones and what genes do they contain?
Is there an alternative assembly for this region and if so, what are the HG reference numbers?

Answers:

1.
The FCGR2C gene has 10 variants in Vega. None of them are protein coding as there is a SNP/DIP in this region of the reference genome that stops the gene from coding and is a known polymorphism and so makes it a polymorphic pseudogene.
Other individuals will have a coding gene, but this cannot be currently represented in the reference genome.

2.
Vega 47 brings up 12 entries. This is a simple text search that looks for the these are also brought up by the search.
There are 2 protein coding gene entries, one on the reference genome and one in a GRC patch region.

In the reference assembly there are 12 alternative variants, 2 of which are protein coding, one is NMD (has a CDS as potentially coding), one transcript and 8 retained introns.

The gene is located on the reverse strand as it is shown below the blue line that represents the BAC genome sequence.

3.
Upstream of this gene are two neighbouring clones AC1091304 and AC138749. There are several pseudogenes here, both processed and unprocessed, plus the GOLGA8F and GOLGA8G genes. This region also has a GRC patch. The HG reference numbers can be viewed in Ensembl, and include HG-753, HG1171, HG923, HG1083 and HG-1022. Details about these regions can be found by clicking on the track.