

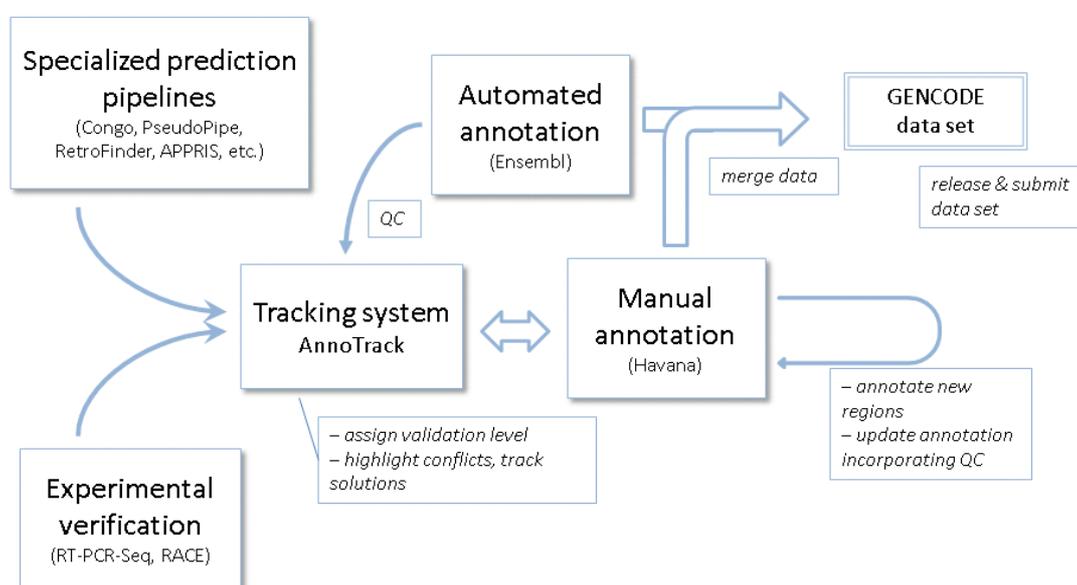
## Overview of the GENCODE reference gene set

### Aims

This module will give an overview of the GENCODE gene set that is available from the genome browsers and 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).

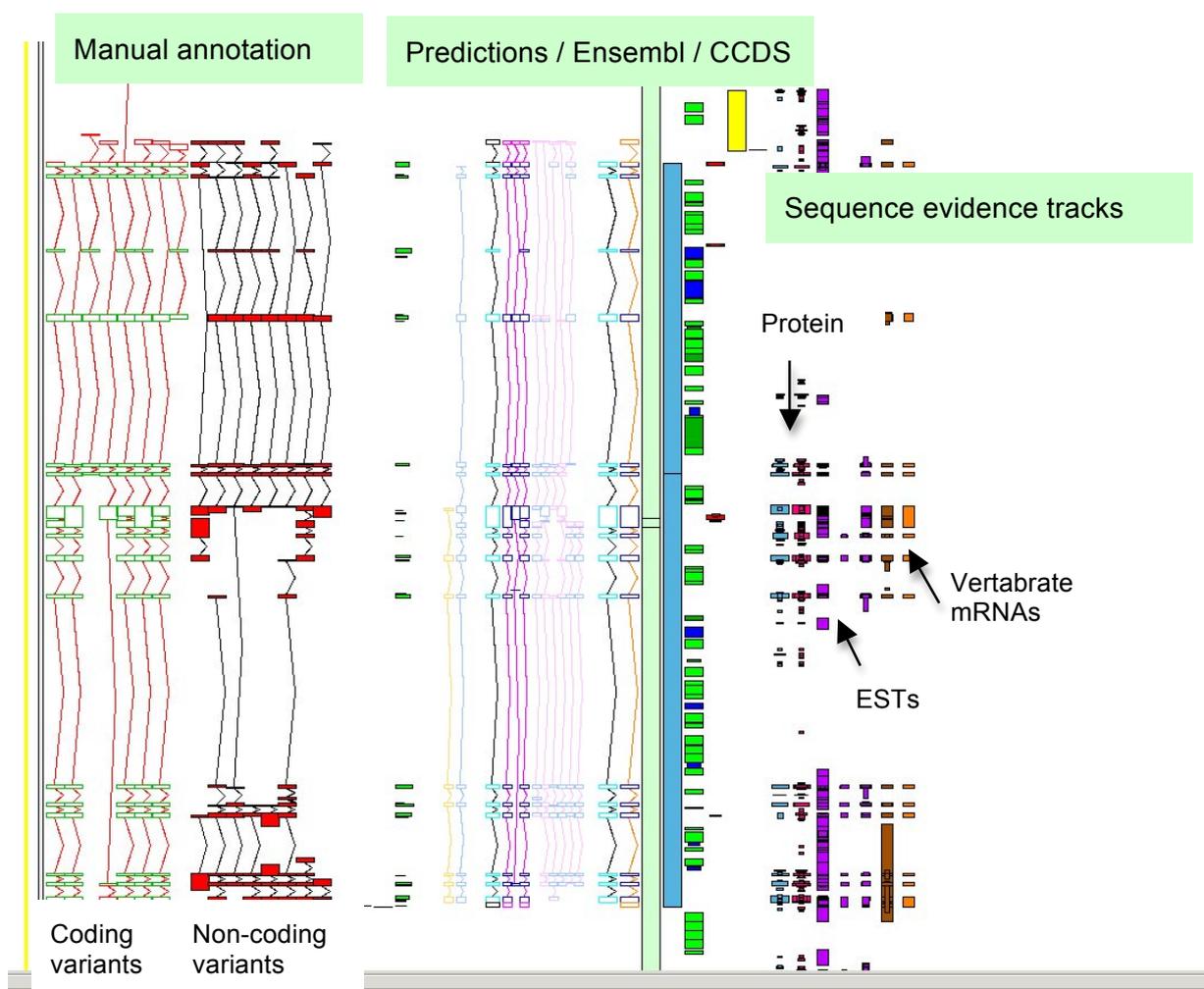
## Manual Genome Annotation

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.

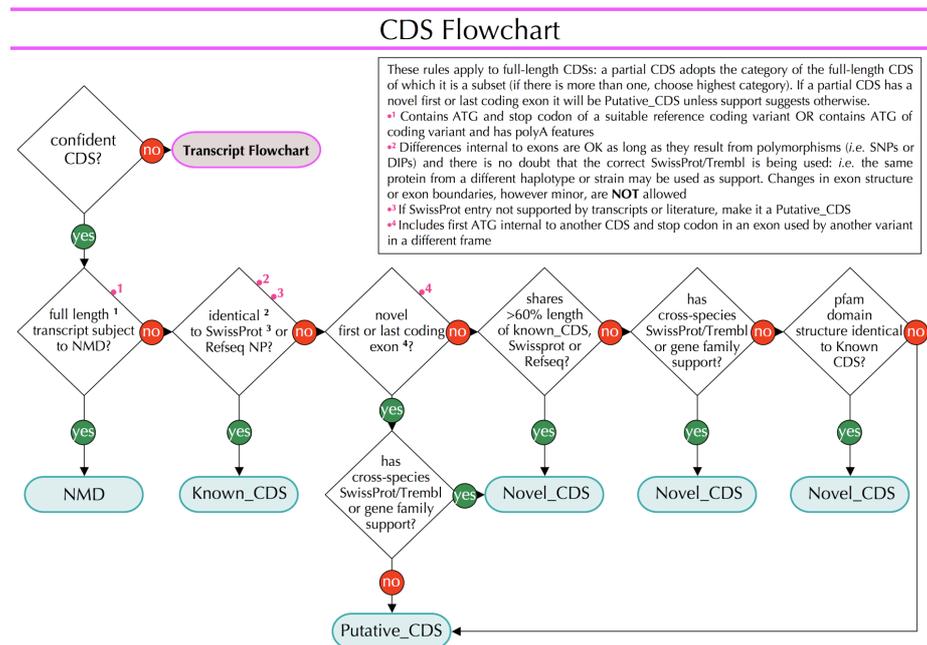
The annotation is then save back to the database. Every few months this data is fed through to Vega and then also incorporated into the Ensembl genebuild. The underlying data for the Vega database is generated by the Havana group. Vega may be browsed and searched in a similar way to Ensembl.

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.

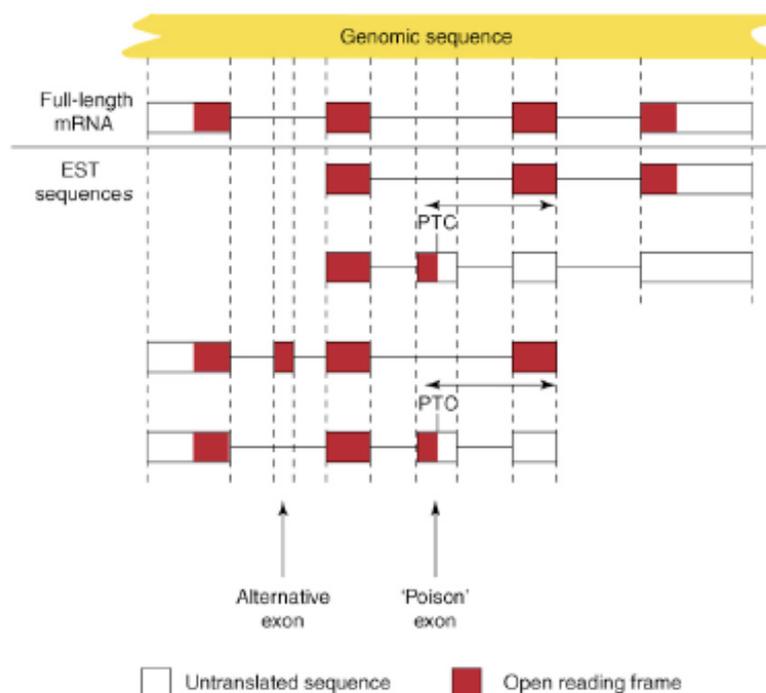
Other columns show Blast hits to DNA and protein databases, repeats and Phastcons regions (evolutionarily conserved regions from 28 vertebrates)



**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 Histocompatibility Complex**

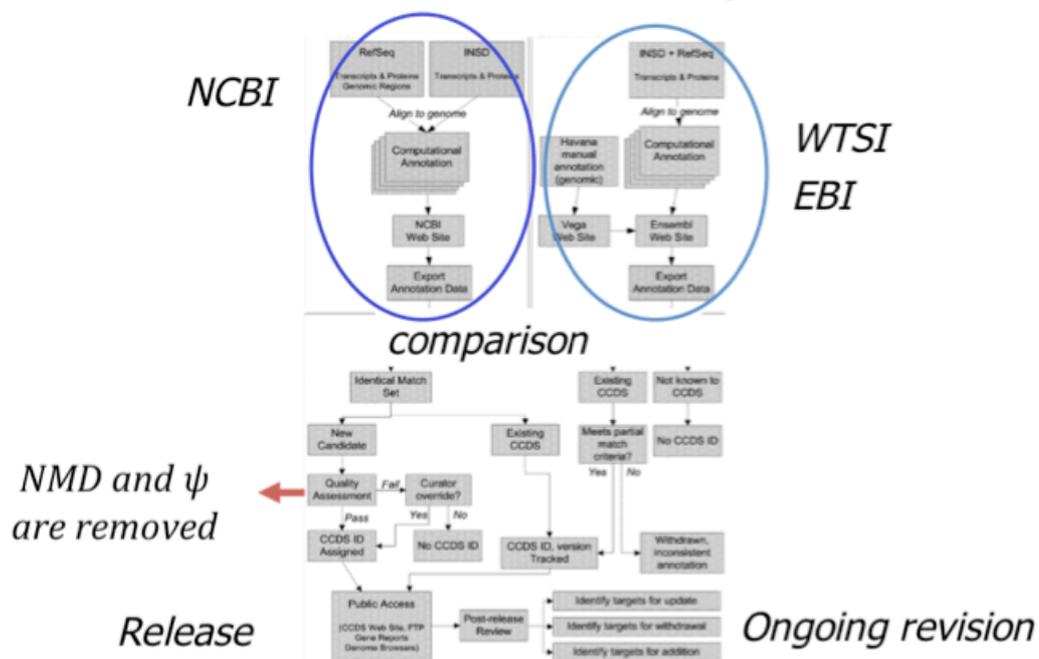
The human major histocompatibility complex (MHC) contains many immune related genes including highly polymorphic examples encoding MHC class I and class II molecules that present antigens to T lymphocytes. Vega has seven human haplotypes of the chromosome 6 MHC region together with reference sequence 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

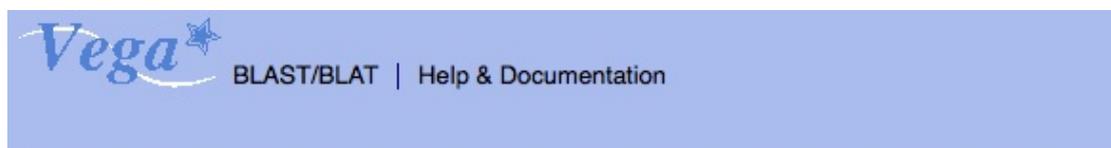
institutes. The long-term goal is to support convergence towards a standard set of gene annotations. The CCDS gene set is generated by Ensembl and NCBI and there is extensive QC by WTSI, NCBI and UCSC. A set of guidelines have been developed for the annotation of coding sequence regions by the collaborating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

## CCDS pipeline: producing consensus



**Worked example 1:**

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



The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality manual annotation of vertebrate finished genome sequence. Human, mouse and zebrafish are in the process of being completely annotated, whereas for other species the annotation is only of specific genomic regions of particular biological interest. The majority of the annotation is from the [HAVANA](#) group at the [Wellcome Trust Sanger Institute](#)



The website is built upon code from the [Ensembl](#) project.

**STEP 1:**  
Load Vega:  
<http://vega.sanger.ac.uk>

**Browse a genome**



**Human** [03-04-2012]  
Ensembl



**Gorilla** [30-03-2009]  
Ensembl



**Zebrafish** [03-04-2012]  
Ensembl



**Wallaby** [30-03-2009]  
Ensembl



**Mouse** [12-01-2012]  
Ensembl



**Pig** [16-05-2007]  
Ensembl



**Chimpanzee** [12-01-2012]  
Ensembl



**Dog** [14-02-2005]  
Ensembl

**STEP 2:**  
Select human genome annotation

**Human (VEGA47)**

About this species

- Description
- Acknowledgements
- What's New
- Sample entry points
  - Chromosome (20)
  - Location (AL035460.15)
  - Gene (MRPS26)
  - Transcript (GNRH2-001)

Configure this page  
Manage your data  
Export data  
Bookmark this page

Search for:

e.g. MRPS26 or AL035460.15

This site presents data from the manual annotation of the human genome, including of the [GRCh37.p5](#) fix and novel patches. The annotation shown in this datafreeze taken on the 19th December 2011 and the gene structures are presented in the merged human geneset shown in Ensembl release 67. This is release 12. All chromosomes apart from 17, 18 and 19 have now undergone a first pass manual annotation by the Havana group at the Wellcome Trust S

Vega also shows manual annotation of loci and regions of particular interest:

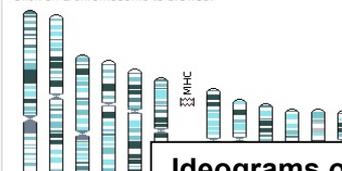
- The [MHC region on chromosome 6](#) in seven haplotypes: COX, QBL, APD, DBB, MANN, MCF, SSTO. These are shown as distinct chromosomes, [6-QBL](#) and are included in the [Vega comparative analysis](#).
- The [LRC region on chromosome 19](#) in nine haplotypes. As for the MHC regions, these are shown as distinct chromosomes, for example [19-COX](#) the [Vega comparative analysis](#).

**STEP 3:**  
Search for gene symbol RECQL4

Genes:	46,369
Processed transcripts	12,165
Pseudogenes	13,362
IG & TR Genes	631
Annotated transcripts	166,571

**Annotation progress**

Click on a chromosome to browse:



**Ideograms of annotated chromosomes and additional information**

Your search of Human with 'RECQL4' returned the following results:

By Feature type		By Species	
Total	1	Total	1
▶ Gene	1	▶ Human	1

By Feature type	
Total	1
▼ Gene	1
Human (1)	

**STEP 4:**  
Expand the Gene section.  
Select the link to the Vega gene

[RECQL4 \[ Havana: OTTHUMG00000165178 \]](#)

**Description** RecQ protein-like 4 [Type: processed transcript Havana]  
**Location** [8:145736671-145743229:-1](#)  
**Source** v47

**STEP 5:**  
List of manually curated transcripts and gene summary.  
Select a transcript by clicking on it.

Notice that there are no protein products for this gene.

Human (VEGA47) Location: 8:145,736,671-145,743,229 Gene: RECQL4

**Gene-based displays**

- Gene summary
- Splice variants (10)
- Supporting evidence
- Sequence
- External references
- Comparative Genomics
  - Genomic alignments
  - Orthologues
  - Alt. alleles
- External Data
  - Personal annotation
  - Other genome browsers
  - Ensembl

Configure this page

Manage your data

Export data

Bookmark this page

Gene: RECQL4 OTTHUMG00000165178

**Description** RecQ protein-like 4  
**Location** [Chromosome 8: 145,736,671-145,743,229](#) reverse strand.  
**Transcripts** This gene has 10 transcripts

Show/hide columns		Filter				
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RECQL4-001	<a href="#">OTTHUMT00000382482</a>	3897	No protein product	-	Processed transcript	-
RECQL4-002	<a href="#">OTTHUMT00000382483</a>	3810	No protein product	-	Processed transcript	-
RECQL4-004	<a href="#">OTTHUMT00000382490</a>	725	No protein product	-	Processed transcript	-
RECQL4-005	<a href="#">OTTHUMT00000382489</a>	686	No protein product	-	Processed transcript	-
RECQL4-007	<a href="#">OTTHUMT00000382488</a>	724	No protein product	-	Processed transcript	-
RECQL4-008	<a href="#">OTTHUMT00000382486</a>	922	No protein product	-	Processed transcript	-
RECQL4-009	<a href="#">OTTHUMT00000382485</a>	324	No protein product	-	Processed transcript	-
RECQL4-011	<a href="#">OTTHUMT00000382484</a>	988	No protein product	-	Processed transcript	-
RECQL4-013	<a href="#">OTTHUMT00000397531</a>	781	No protein product	-	Processed transcript	-
RECQL4-003	<a href="#">OTTHUMT00000382491</a>	517	No protein product	-	Retained intron	-

Gene summary [help](#)

**Curated Locus** [RECQL4](#) (HGNC Symbol)  
**Synonyms** RecQ4 [To view all genes linked to the name [click here](#).]  
**Gene type** Known processed transcript [[Definition](#)]  
**Author** This gene was annotated by Havana <[vega@sanger.ac.uk](mailto:vega@sanger.ac.uk)>  
**Version & date** Version 2, last modified on 08/10/2010 (Created on 22/04/2010)  
**Other assemblies** This gene maps to to [145,736,671-145,743,229](#) in GRCh37 (Ensembl) coordinates. [Jump](#) to this stable ID in Ensembl  
**Curation Method** See this [description](#) of the manual annotation process  
**Alternative genes** **Ensembl gene:** [ENSG00000160957](#) [[view all locations](#)]

**Transcript summary** [help](#)



[Export Image](#)

<b>Statistics</b>	<b>Exons:</b> 20 <b>Transcript length:</b> 3,897 bps
<b>Class</b>	processed transcript ( <a href="#">Definition</a> )
<b>Author</b>	This transcript was annotated by Havana
<b>Version &amp; date</b>	Version 1, last modified on 08/10/2010 (Created on 22/04/2010)
<b>Alternative symbols</b>	CTD-2517M22.13-001
<b>Remarks</b>	suspected genomic sequence error affecting CDS in exon 14
<b>Other assemblies</b>	This transcript maps to <a href="#">145,736,671-145,743,229</a> in GRCh37 (Ensembl) coordinates. <a href="#">Jump</a> to this stable ID in Ensembl
<b>Alternative transcripts</b>	<b>Ensembl transcript having exact match with Havana:</b> <a href="#">ENST00000532237</a> ( <a href="#">view all locations</a> )
<b>Curation Method</b>	See this <a href="#">description</a> of the manual annotation process

**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.

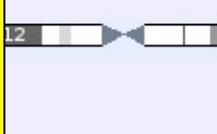
**Vega** BLAST/BLAT | [Help & Documentation](#)

Human (VEGA47) ▾ **Location:** 8:145,736,671-145,743,229 **Gene:** RECQL4 **Transcript:** RECQL4-001

**Location-based displays**

- Whole genome
- Chromosome summary
- Region overview
- **Region in detail**
- Comparative Genomics
  - Alignments (image)
  - Alignments (text)
  - Multi slice view
- Markers
- Other genome browsers
  - [Ensembl](#)

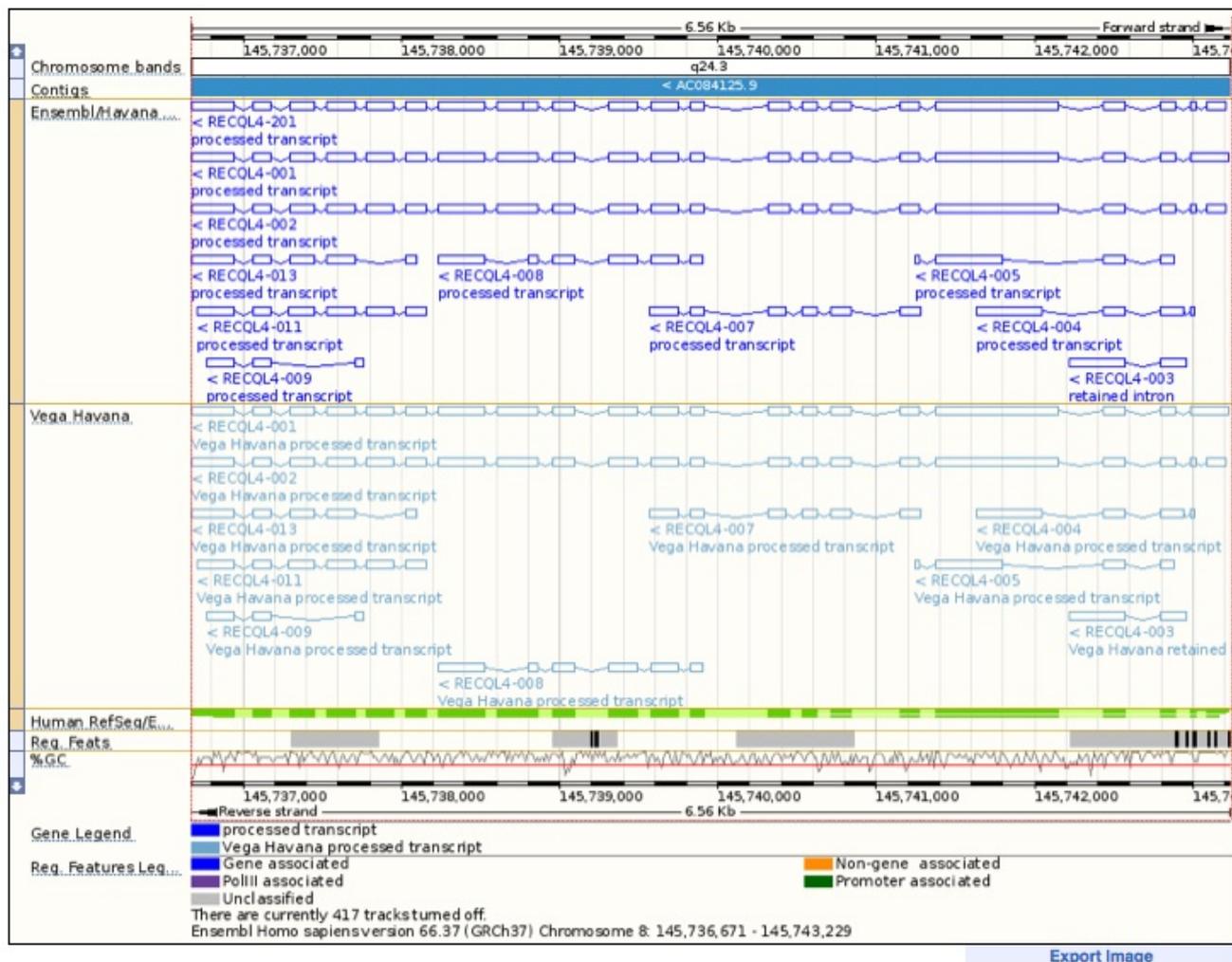
**Chromosome 8:**



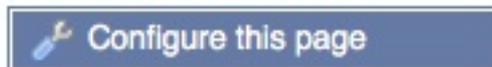
**STEP 7:**  
Click on the location tab at the top of the page, then click on the side link to Ensembl.

**Region i**

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:

**Sequence and assembly**

Enable/disable all Misc. regions & clones

- VEGA46 assembly
- Vega clones
- 1Mb clone set
- 30k clone set
- 32k clone set
- ENCODE excluded regions
- Encode regions
- Genomic contigs
- Tilepath

External data (DAS)

- DAS CHORI17 BACs
- DAS CPG island clones
- DAS GRC region NCBI\_37
- DAS Pig BAC ends

**STEP 8:**  
GRC track in under external data (DAS)

**Key**

- Track style
- Forward strand
- Reverse strand
- Favourite track
- Track information

**External tracks**

- DAS Distributed Annotation Source
- Temp Custom track - uploaded data
- URL Custom track - UCSC-style web resource
- Saved Custom data saved to your user account

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

**STEP 9:**  
Click on GRC track to give information about the genomic error.

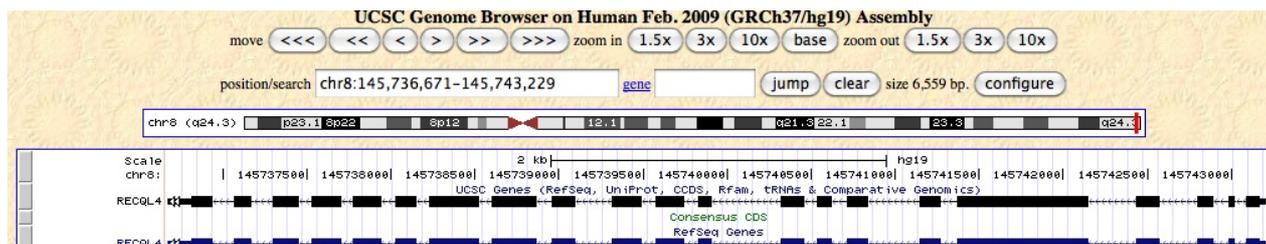
**GRC report for HG-334**

There is a possible functional difference between the proteins encoded by RefSeq NM\_004260.2 (RECQL4) and the genomic region to which it aligns.

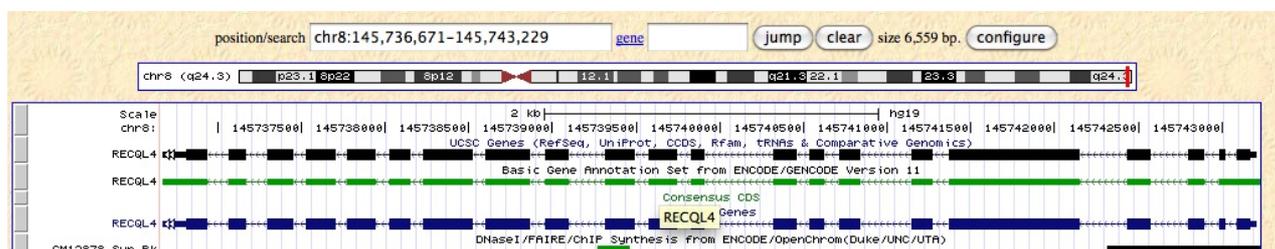
We can now jump to the same gene in the UCSC genome browser, by clicking on the UCSC link in the side bar



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)		
<b>GENCODE Transcript Annotation ENST00000532237.1 (RECQL4)</b>		
	Transcript	Gene
Gencode id	<a href="#">ENST00000532237.1</a>	<a href="#">ENSG00000160957.7</a>
HAVANA manual id	<a href="#">OTTHUMT00000382482.1</a>	<a href="#">OTTHUMG00000165178.2</a>
Position	<a href="#">chr8:145736671-145743229</a>	<a href="#">chr8:145736671-145743229</a>
Strand	-	
<u>Biotype</u>	processed_transcript	processed_transcript
Status	KNOWN	KNOWN
Annotation Level	manual (2)	
Annotation Method	manual	manual & automatic
<u>Transcription Support Level</u>	<a href="#">ts1</a>	
HUGO gene	RECQL4	
CCDS		
Tags		
Sequences		
<a href="#">Predicted mRNA</a>		
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.

2 Genes match your query ('ABO') in Human

[ABO](#) [ Havana: OTTHUMG00000020872 ]

**Description** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) [Type: processed transcript Havana]

**Location** [9:136131053-136150617:-1](#)

**Source** v47

[ABO](#) [ Havana: OTTHUMG00000174691 ]

**Description** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) [Type: protein coding Havana]

**Location** [HG79\\_PATCH:136131200-136150736:-1](#)

**Source** v47

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

**Gene: ABO OTTHUMG00000020872**

**Description** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase),

**Location** [Chromosome 9: 136,131,053-136,150,617](#) reverse strand.

**Transcripts** This gene has 1 transcript

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
ABO-001	<a href="#">OTTHUMT00000054907</a>	1076	No protein product	-	Processed transcript	-

**Gene summary help**

**Curated Locus** [ABO](#) (HGNC Symbol)

**Synonyms** A3GALNT, A3GALT1 [To view all genes linked to the name [click here](#).

**Gene type** Known processed transcript [Definition]

**Author** This gene was annotated by Havana <[vega@sanger.ac.uk](mailto:vega@sanger.ac.uk)>

**Version & date** Version 3, last modified on 14/09/2011 (Created on 11/12/2003)

**Other assemblies** This gene maps to to [136,131,053-136,150,617](#) in GRCh37 (Ensembl [Jump](#) to this stable ID in Ensembl)

**Curation Method** See this [description](#) of the manual annotation process

**Alternative genes** Ensembl gene: [ENSG00000175164](#) [view all locations]

**Contigs** [AL792364.9.1.87903](#) > [AL158826.23.1.184513](#) >

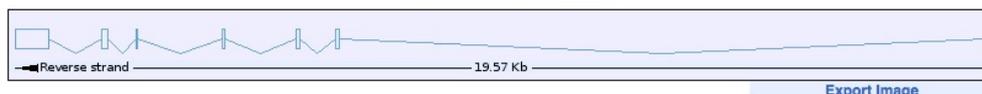
**Havana gene.** < ABO-001 Havana processed transcript

The gene is a non-coding transcript.

## Transcript: ABO-001 OTTHUMT00000054907

**Description** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)  
**Location** [Chromosome 9: 136,131,053-136,150,617](#) reverse strand.  
**Gene**  This transcript is a product of gene [OTTHUMG00000020872](#) - This gene has 1 transcript

Show/hide columns		Filter				
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
ABO-001	<a href="#">OTTHUMT00000054907</a>	1076	No protein product	-	Processed transcript	-

Transcript summary [help](#)

**Statistics** Exons: 7 Transcript length: 1,076 bps  
**Class** processed transcript [\[Definition\]](#)  
**Author** This transcript was annotated by Havana  
**Version & date** Version 3, last modified on 14/09/2011 (Created on 11/12/2003)  
**Alternative symbols** RP11-430N14.3-001  
**Remarks** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase), ABO-001 allele  
 ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase), ABO-002 allele  
 The ABO gene in this individual produces a truncated protein without functional glycosyltransferase activity indicative of blood group O  
**Other assemblies** This transcript maps to to [136,131,053-136,150,617](#) in GRCh37 (Ensembl) coordinates.  
[Jump](#) to this stable ID in Ensembl  
**Alternative transcripts** **Ensembl transcript having exact match with Havana:** [ENST00000453660](#) [\[view all locations\]](#)  
**Curation Method** See this [description](#) of the manual annotation process

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.

[ABO \[ Havana: OTTHUMG00000174691 \]](#)

**Description** ABO blood group (transferase A,  
**Location** [HG79\\_PATCH:136131200-136131200](#)  
**Source** v47

**STEP 3:**  
 Click onto the gene ID for the HG\_79 PATCH entry.

The gene is now protein coding in the patch assembly:

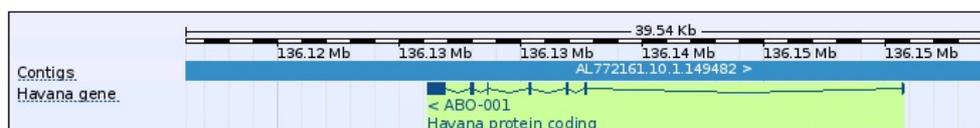
**Gene: ABO OTTHUMG00000174691**

**Description** ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)  
**Location** [Chromosome HG79 PATCH: 136,131,200-136,150,736](#) reverse strand.  
**Transcripts**  This gene has 1 transcript

Show/hide columns		Filter				
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
ABO-001	<a href="#">OTTHUMT00000427219</a>	1077	<a href="#">OTTHUMP00000253927</a>	354	Protein coding	-

**Gene summary** [help](#)

**Curated Locus** [ABO](#) (HGNC Symbol)  
**Synonyms** A3GALNT, A3GALT1 [To view all genes linked to the name [click here](#).]  
**Gene type** Known protein coding [[Definition](#)]  
**Author** This gene was annotated by Havana <[vega@sanger.ac.uk](mailto:vega@sanger.ac.uk)>  
**Version & date** Version 2, last modified on 17/08/2011 (Created on 11/08/2011)  
**Other assemblies** This gene maps to to [136,131,200-136,150,736](#) in GRCh37 (Ensembl) coordinates. [Jump](#) to this stable ID in Ensembl  
**Curation Method** See this [description](#) of the manual annotation process  
**Alternative genes** **Ensembl gene:** [ENSG00000256062](#) [[view all locations](#)]



Patch assembly in Ensembl:

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

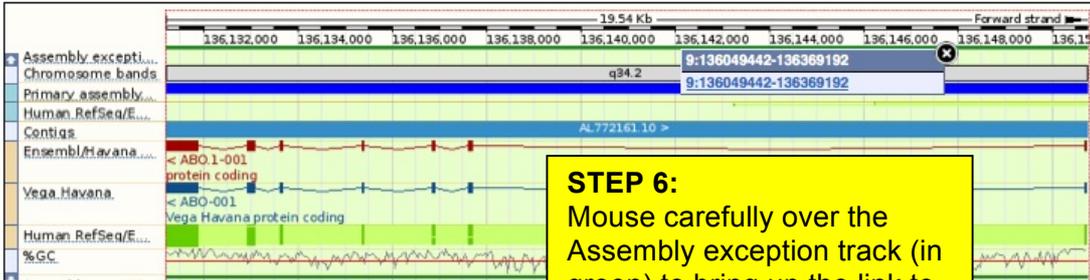
**Warning/Error(s):**

- Sorry, couldn't locate ChrHG79\_PATCH:136131200-136150736 in genome database

OK

Location:

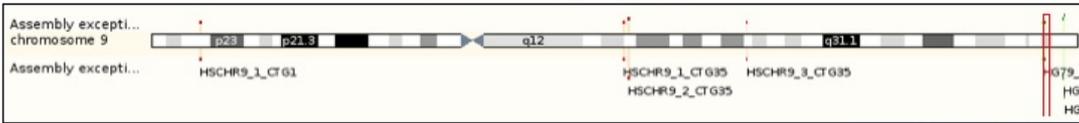
<< < + | - > >>



**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.

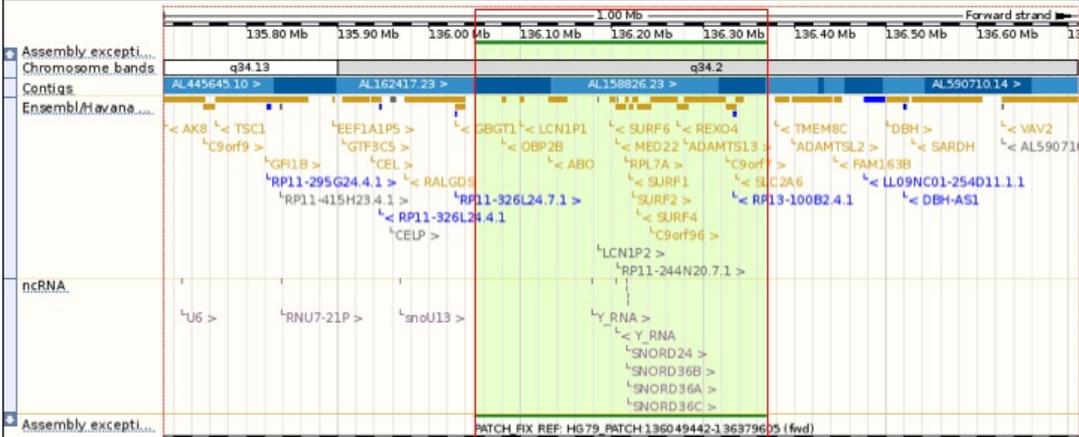
We are now back in reference human genome assembly.

Chromosome 9: 136,049,442-136,369,192

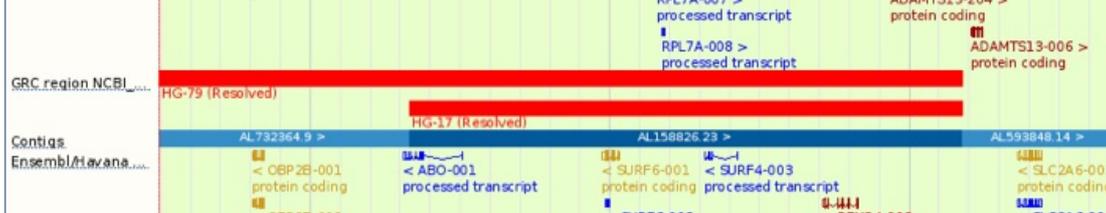


[Export image](#)

Region in detail [help](#)



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

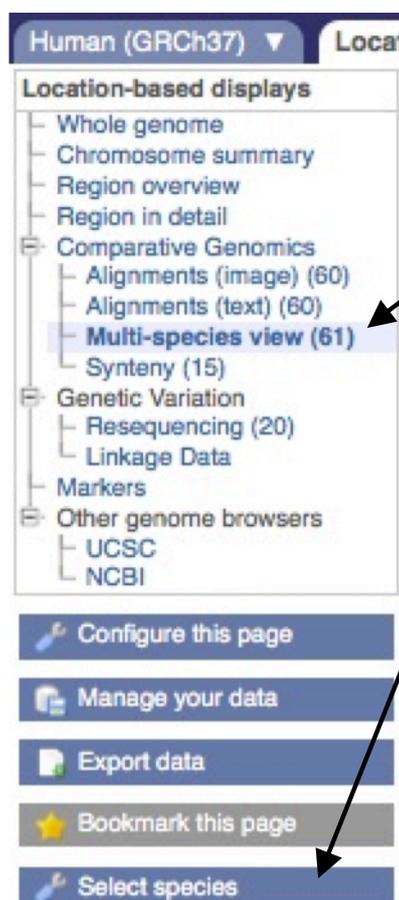
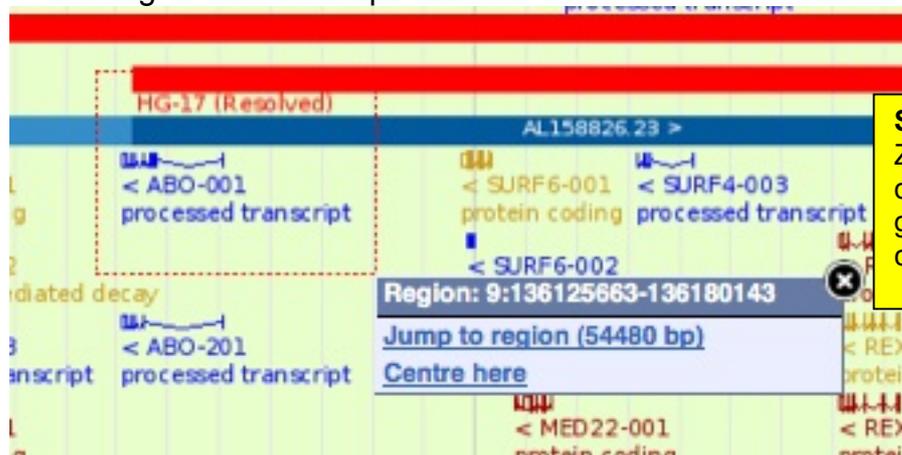


GRC region NCBI... HG-79 (Resolved) HG-17 (Resolved)

Contigs AL732364.9 AL158826.23 AL593848.14

Ensembl/Havana... < OBP2B-001 protein coding < ABO-001 processed transcript < SURF6-001 protein coding < SURF4-003 processed transcript < SLC2A6-002 protein coding < OBP2B-002 < SURF6-002 < REXO4-002 < SLC2A6-003

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



Configure Multi-species Overview | Select species | Manage Configurations | Custom Data

**Selected species**

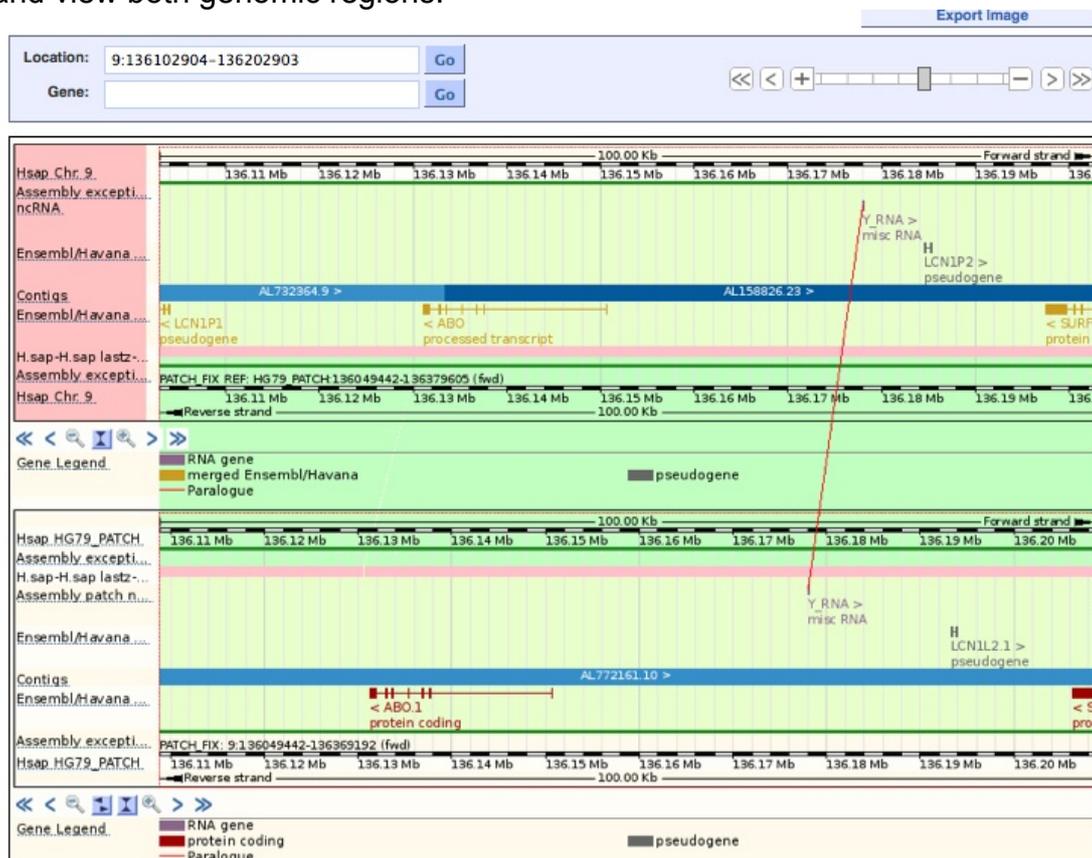
- Human (Homo sapiens) HG79\_PATCH - lastz patch

**Unselected species**

- Alpaca (Vicugna pacos) - blastz
- Anole Lizard (Anolis carolinensis) - translated blat
- Armadillo (Dasypus novemcinctus) - blastz
- Bushbaby (Otolemur garnettii) - lastz
- Cat (Felis catus) - blastz

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.

