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).

1

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 nonsensemediated 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.



CCDS pipeline: producing consensus

Worked example 1:

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



(

V Feature type			By Specie	06			
Total		1	Total	63		1	
Gono		1				1	
			p numa				
					STED A		
By Feature type					Expand	r. the Gene	
fotal		1			section		
Gene		1			Select t	he link to the	<u> </u>
Human (1)					Vega g	ene	
RECOLA [Hava	na: OTTHUM	4600001651	781				
Description	Rec	Q protein-like	4 [Type: p	rocessed tra	nscript Ha	avana]	
Location	8:14	45736671-145	743229:-1				
Course	0.14	10100071-140	TOLLO.1			STEP 5:	
Source	V47					List of ma	nually
						curated tra	anscripts
						and gene	summary.
otice that there	e are no pro	otein produc	ts for this	s gene.		Select a tr	summary. ranscript b
otice that there	e are no pro	otein produc	ts for this	s gene.		and gene Select a tr	summary. ranscript b n it
otice that there	e are no pro	otein produc	ts for this	s gene.		and gene Select a tr clicking or	summary. ranscript b n it.
otice that there	e are no pro	otein produc	ets for this	s gene.		and gene Select a tr clicking or	summary. ranscript b n it.
otice that there	e are no pro	otein produc	ts for this	s gene.		and gene Select a tr clicking or	summary. ranscript b n it.
otice that there an (VEGA47) V Location based displays the summary	e are no pro	otein produc	ets for this	s gene.	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b n it.
n (VEGA47) ▼ Locati based displays re summary ce variants (10) poporting evidence	e are no pro	otein produc	ets for this	s gene.	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b n it.
n (VEGA47) ▼ Locati based displays te summary ce variants (10) porting evidence juence page lefaceopo	e are no pro	otein product	ets for this	S gene. Gene: FEC	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b n it.
n (VEGA47) V Location Loc	e are no pro	otein product 145,743,229 Gene: RecQ protein-like 4 <u>Chromosome 8: 145,73</u> This gene has 10 transc	Ets for this RECQL4 6.671-145.743.22 cripts	S gene. Gene: TEC 29 reverse strand.	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b
n (VEGA47) V Locati ased displays be summary ce variants (10) porting evidence uence mal references parative Genomics ienomic alignments irthologues	ion: 8:145,736,671-1 Description Location Transcripts Show/hide	otein product 145,743,229 Gene: RecQ protein-like 4 <u>Chromosome 8: 145,73</u> This gene has 10 transc columns	ets for this RECQL4 6.671-145.743.22 pripts	S gene. Gene: FEC	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b n it.
n (VEGA47) Locati ased displays te summary ce variants (10) porting evidence uence uence trial ferences that displays that displays	ion: 8:145,736,671-1 Description Location Transcripts Show/hide Name	otein product	RECQL4	S gene. Gene: FEC 29 reverse strand.	CQL4 OTTHUMG	and gene Select a tr clicking or	summary. ranscript b n it.
n (VEGA47) Locati ased displays e summary be variants (10) porting evidence uence uence uence uence intral leferences the alignments intrologues the align	ion: 8:145,736,671-1 Description Location Transcripts Show/hide RECQL4-001	otein product	RECOL4	S GENE. Gene: FEC 29 reverse strand. Protein ID	CQL4 OTTHUMG	and gene Select a tr clicking or 200000165178	summary. ranscript b n it.
n (VEGA47) ▼ Locati based displays e summary be variants (10) porting evidence uence uence uence trinal references parative Genomics enomic alignments tritologues trit. alleles trital Data ersonal annotation er genome browsers nsembl	ion: 8:145,736,671-1 Description Location Transcripts	otein product	ets for this RECOL4 6,671-145,743,222 aripts	S Gene: FEC Gene: FEC 29 reverse strand.	Length (aa)	And gene Select a tr clicking or 00000165178	summary. ranscript b n it.
tice that there (VEGA47) ▼ Locati ased displays e summary be variants (10) oorting evidence uence mal references uence mal references uence thologues th. alleles mal Data ersonal annotation arg genome browsers nsembl	e are no pro	otein product	Image: second	S Gene: FEC	Length (aa)	And gene Select a tr clicking or clicking	summary. ranscript b n it.
tice that there (VEGA47) ▼ Location Locatio	e are no pro	otein product	Image: state state state Image: state state 86.671-145.743.222 bright (bp) 2 3897 2 3897 2 3897 2 3897 2 3897 3 3810 725 686 3	S Gene: FEC	Length (aa) - - -	and gene Select a tr clicking or clicking	summary. ranscript b n it.
n (VEGA47) ▼ Locati ased displays the summary ce variants (10) porting evidence uence immal references inparative Genomics tenomics alignments inthologues it. alleles errand pata ersonal annotation ar genome browsers nsembl ponfigure this page anage your data	e are no pro	otein product	RECQL4 6,671-145,743,222 bright (bp) € 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 3 3810 2 2 686 2 2 2 382	S Gene: FEC Gene: FEC 20 reverse strand.	Length (aa) 🔶	and gene Select a tr clicking or clicking or clicking or boothing clicking or clicking or	summary. ranscript b n it.
n (VEGA47) ▼ Locati ased displays e summary be variants (10) porting evidence uence irmal references inparative Genomics enomic alignments inthologues it, alleles irmal Data ersonal annotation er genome browsers insembl onfigure this page anage your data	e are no pro	otein product	Ets for this RECQL4 6.671-145,743,222 rripts 2 3897 3 3810 2 2 686 3 2 922 5 324	S Gene: FEC Gene: FEC 20 reverse strand.	Length (aa) 🔶 - - - - - - -	and gene Select a tr clicking or clicking or clicking or boother clicking or clicking or c	summary. ranscript b n it.
n (VEGA47) ▼ Locati ased displays e summary e variants (10) porting evidence uence mail references upparative Genomics enomic alignments inthologues tt. alleles irmal Data ersonal annotation er genome browsers nsembl onfigure this page anage your data	e are no pro	otein product	Ets for this RECQL4 6.671-145,743,222 rripts 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 324 4 988	S Gene: Fec Gene: Fec 29 reverse strand.	Length (aa) € - - - - - - - - - -	And gene Select a tr clicking or clicking	summary. ranscript b n it. CCDS ÷ - - - - - - - - - - - - - - - - - - -
tice that there (VEGA47) ▼ Locati ased displays e summary the variants (10) booting evidence Jence mai references parative Genomics enomic alignments thologues t. alleles mai Data ersonal annotation er genome browsers nsembl onfigure this page anage your data port data bokmark this page	e are no pro	Transcript ID 0THUMT00000382482 0TTHUMT00000382482 0TTHUMT0000382482 0TTHUMT00000382482 0TTHUMT000000382482	Image: start start Image: start start start 6.671-145,743,222 rripts 1 2 3 3810 2 3810 2 2 3810 2 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3810 2 3824 4 988 2 388	S Gene: Fee Gene: Fee 29 reverse strand.	Length (aa) ∳ - - - - - - - - - - - - -	And gene Select a tr clicking or clicking	summary. ranscript b n it.
tice that there (VEGA47) ▼ Locati ased displays e summary te variants (10) booting evidence Jence mal references parative Genomics enomic alignments thologues t. alleles mai Data arsonal annotation or genome browsers tsembl Infigure this page port data lokmark this page	e are no pro	Transcript ID OTTHUMT00000382482 OTTHUMT00000382482 OTTHUMT0000382482 OTTHUMT00000382482	Image: start start Image: start start start Back start s	S Gene: Fee Gene: Fee 29 reverse strand.	Length (aa) € - - - - - - - - - - - - - - - - - - -	and gene Select a tr Clicking or Clicking	summary. ranscript b n it.
tice that there (VEGA47) ▼ Locati ased displays e summary e variants (10) porting evidence uence uence enomic alignments thologues tt. alleles rmal Data ersonal annotation er genome browsers nsembl unfigure this page anage your data port data pokmark this page	e are no pro	Transcript ID OTHUMT00000382482 OTHUMT00000382482 OTTHUMT00000382482	Image Image 6.671-145.743.222 bright (bp) 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 2 3897 3 2 324 988 1 517	S Gene: PEC Gene: PEC 29 reverse strand.	Length (aa) 	And gene Select a tr clicking or clicking	summary. ranscript b n it.
Dice that there n (VEGA47) ▼ Location based displays the summary ce variants (10) porting evidence uence semal references nparative Genomics senomic alignments borthologues tit. alleles semal Data tersonal annotation er genome browsers insembl configure this page anage your data xport data bokmark this page	e are no pro	Transcript ID OTHUMT00000382482 OTTHUMT00000382482	ets for this RECQL4 6,671-145,743,222 pripts 4 0806 3 725 2 0806 3 724 3 922 2 324 4 988 1 781 1 517	S Gene: rec Gene: rec Protein ID No protein product No protein product	Length (aa)	And gene Select a tr Clicking or Clicking	summary. ranscript b n it.
Dice that there n (VEGA47) V Locations based displays be summary commar	e are no pro	Transcript ID OTHUMT00000382482 OTTHUMT00000382482	ets for this RECQL4 6,671-145,743,222 pripts 2 3897 3 3810 2 725 2 3897 3 3810 2 725 2 686 3 724 2 922 2 324 4 988 1 781 1 517 ymbol)	S Gene: rec Gene: rec Protein ID No protein product No protein product	Length (aa)	And gene Select a tr Clicking or Clicking	summary. ranscript b n it.
btice that there n (VEGA47) ▼ Locati hased displays le summary ce variants (10) porting evidence uence trial references parative Genomics ternomic alignments brithologues It. alleles mail Data tersonal annotation ar genome browsers nsembl brithologues tt. alleles mail pata tersonal annotation ar genome browsers nsembl brithologues tt. alleles trial pata tersonal annotation ar genome browsers nsembl brithologues tt. alleles trial pata tersonal annotation ar genome browsers nsembl brithologues tt. alleles trial pata tersonal annotation tersonal annotatio	e are no pro	Transcript ID OTHUMT00000382482 OTHUMT00000382482 OTTHUMT00000382482 OTTHUMT00000382484 OTTHUMT00000382482 OTTHUMT00000382484	Ets for this RECQL4 6,671-145,743,222 pripts 4 0866 3 724 2 0866 3 724 2 0866 3 724 4 088 1 781 1 517 ymbol) genes linked to th	S Gene: rec Gene: rec Protein ID No protein product No protein product	Length (aa)	And gene Select a tr clicking or clicking	summary. ranscript b n it.
btice that there n (VEGA47) ▼ Location based displays the summary ce variants (10) porting evidence uence email references nparative Genomics tenomic alignments brithologues tt. alleles email Data tersonal annotation er genome browsers insembl configure this page anage your data pokmark this page	e are no pro	Transcript ID OTHUMT00000382482 OTHUMT00000382482 OTTHUMT00000382482 OTTHUMT00000382484 OTHUMT00000382484 OTHUMT00000382484 OTTHUMT00000382484 OTTHUMT00000382484 OTTHUMT00000382484 <	Image: starting start	S Gene: rec Gene: rec Protein ID No protein product No protein product	Length (aa)	And gene Select a tr clicking or clicking	summary. ranscript b n it.
btice that there of (VEGA47) ▼ Location based displays the summary the summary	e are no pro	Transcript ID OTTHUMT00000382482	ets for this RECQL4 6,671-145,743,222 cripts	S Gene: rec Gene: rec Protein ID No protein product No protein	Length (aa)	and gene Select a tr clicking or clicking	summary. ranscript b n it.
otice that there an (VEGA47) ▼ Locati based displays ne summary lice variants (10) sporting evidence quence ernal references mparative Genomics Senomic alignments Orthologues Alt. alleles ernal Data Personal annotation ler genome browsers Ensembl configure this page tanage your data xport data cookmark this page	e are no pro	Otein product 145,743,229 Gene: 145,743,229 Gene: RecQ protein-like 4 Chromosome 8: 145,73 This gene has 10 transc Columns Transcript ID OTTHUMT00000382482 OTTHUMT00000382482 OTTHUMT00000382482 OTTHUMT00000382482<	Image: sets for this RECOL4 6,671-145,743,222 cripts 1 1 2 3897 3 3810 2 2 3897 3 3810 2 2 3897 3 3810 2 3897 3 2 3897 3 3810 2 324 4 988 1 517 ymbol) genes linked to th ranscript [Definitic otated by Havana ified on 08/10/201	S Gene: Rec Gene: Rec Protein ID No protein product No protein	Length (aa) (*)	And gene Select a tr clicking or clicking	summary. ranscript b n it.
an (VEGA47) ▼ Locati based displays ne summary loce variants (10) oporting evidence quence emal references mparative Genomics Darative Genomics Darative Genomics Parative Genomics Darative Genomics Parative Genomics Parative Genomics Data Personal annotation ler genome browsers Ensembl configure this page tanage your data	e are no pro	otein product	Image: sets for this RECOL4 6,671-145,743,222 cripts 1 1 2 3897 3 3810 2 2 3897 3 3810 2 2 3897 3 3810 2 2 3897 3 3810 2 3824 4 988 1 517 ymbol) genes linked to th ranscript [Definitic otated by Havana ified on 08/10/201 to 145,736,671-11	S Gene: rec Gene: rec Protein ID No protein product No protein	Length (aa) (*) - - - - - - - - - - - - - - - - - - -	and gene Select a tr clicking or clicking	summary. ranscript b n it.
otice that there an (VEGA47) ▼ Locati based displays ne summary lice variants (10) oporting evidence quence emal references mparative Genomics Data Personal annotation her genome browsers Ensembl configure this page tanage your data Export data bookmark this page	e are no pro	otein product I45,743,229 Gene: I45,743,229 Gene: RecQ protein-like 4 Chromosome 8: 145,73 Transcript ID OTTHUMT00000382482 OTTHUMT0	Image: starting start Image: starting starti	S Gene: rec Gene: rec Protein ID No protein product No protein	Length (aa)	and gene Select a tr clicking or clicking	summary. ranscript b n it.
an (VEGA47) ▼ Locati based displays ne summary lice variants (10) oporting evidence quence ernal references mparative Genomics Genomic alignments Orthologues Att. alleles ernal Data Personal annotation her genome browsers Ensembl Xonfigure this page Anaage your data Export data	e are no pro	Otein product 145,743,229 Gene: 145,743,229 Gene: RecQ protein-like 4 Chromosome 8: 145,73 This gene has 10 transc Columns Transcript ID OTTHUMT00000382482 OTTHUMT00000382482 OTTHUMT000000382482 OTTHUMT000000	Image: starting start BECQL4 6,671-145,743,222 cripts 1 1 2 3897 3 3810 2 3897 324 988 781 517 ymbol) genes linked to th ranscript [Definitic on 08/10/201 to 145,736,671-11 ID in Ensembli a of the manual ar Encoccccccccccccccccccccccccccccccccccc	S Gene: rec Gene: rec Protein ID No protein product No protein	Length (aa)	and gene Select a tr clicking or clicking	summary. ranscript b n it.

		Transcript summary he!p	
	F	Accesse strand 6.56 Kb	Export Image
	Statistics Class	Exons: 20 Transcript length: 3,897 bps processed transcript [Definition]	STEP 6:
	Author Version & date	This transcript was annotated by Havana Version 1, last modified on 08/10/2010 (Created on 22/04/2010)	explains that there is a
<	Remarks Other assemblies	suspected genomic sequence error affecting CDS in exon 14 This transcript maps to to 145 736 671 145 746 229 in GRCh37 (Ensembl) coordinates.	region that affects the
	Alternative transcript	<u>Jump</u> to this stable ID in Ensembl Ensembl transcript having exact match with Havana: ENST00000532237 [view all locations]	CDS.
	Curation Method	See this description of the manual annotation process	

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* BLAS	T/BLAT Help & Docu	mentation			
Human (VEGA47) 🔻	Location: 8:145,736,6	71-145,743,229	Gene: RECQL4	Transcri	pt: RECQL4-001
Location-based displays					Chromosome 8:
Whole genome Chromosome summary Region overview Region in detail Comparative Genomics Alignments (image) Alignments (text) Multi slice view Markers		Assembly except chromosome 8 Assembly except	STEP 7: Click on the loca tab at the top of page, then click side link to Ense	ation the on the embl.	
					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

Configure this page

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:

Configure Region Image	onfigure Overv	view Imag	ge Mar	nage Conf	iguratio	ns Custom Data
Active tracks	Se	quen	ce and a	assem	bly	
Favourite tracks	Eng	Enchle/dischle all Miss regions & clones				
Track order		ibic/uia	able all Mia	se, regio		ones
Search results		VEGA46	6 assembly			
E Sequence and assembly	(2/24)	Vega clo	nes			
- Sequence	(1/4)	1Mb clor	ne set			
 Markers Dites features 	(0/1)	30k clon	e set			
- Simple features	(0/2)	32k clon	e set		r	
 Misc. regions & clones 	(1/13)	ENCOD	E excluded re	nions		STEP 8:
Genes and transcripts	(11/48)	Encode	regione	giono		GRC track in under
- Genes	(11/14)	Cases	regiona			external data (DAS)
 Prediction transcripts BNA-Seg models 	(0/2)	Genomic	c contigs			
mPNA and protein alignments	(1/15)	Tilepath				
mRNA alignments	(1/3) Exte	ernal dat	a (DAS)		-	
 EST alignments 	(0/1)	DAS	CHORI17 BA	ACs		
 Protein alignments Protein features 	(0/5)	CPG island clones				
ac BNA	(0/2)	CO DAS	GBC region	NCBL 37		
Inchina	(U/3) E2	C DAS	Die DAG and	NODI_37		
Other DNA alignments	(1/1)	ODAS	Pig BAC end	IS		
Germline variation	(0/78)			Esterne	I treaks	1
 DOSNP 1000 Genomes & HapMap 	(0/20)			Externa	I Udcks	
- Phenotype and curated variants	(0/10)	Track sty	/le	🙆 DAS	Distribu	ited Annotation Source
 Individual genomes 	(0/14)	Forward	strand	C. Temp	Custom	n track - uploaded data
 Arrays and other Failed variants 	(0/13)	Reverse	strand	GURI	Custom	track - UCSC-style web resource
 Structural variants 	(0/11)	neverse	Straniu	Conc.	Custom	
Somatic mutations	(0/44)	Favourite	e track	Saved	Custom	ata saveo to your user account
Regulation	(1/113)	Track inf	ormation			
Regulatory features	(1/20)					

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

			6.56 Kb	Forward strand
Chromosome bands	145.737.000	145,738,000 145,739,000	145.740.000 J	45,741,000 145,742,000 145
GRC region NCBL	HG-334 (Awaiting Exp	HG-334 (Awaiting Expti Data)	8	
Contigs Ensembl/Havana	< RECQL4-201 processed transcript	Type: Clone Problem:There is a possible functional difference between the proteins encoded by RefSeq NM_004260.2		-a-ca-a.a.
	Processed transcript < RECQL4-002 processed transcript	(RECQL4) and the genomic region to which it aligns. Method: Clone Problem; Status Awaiting Exptl Data		Click on GRC track to
	< RECQL4-013 processed transcript < RECQL4-011	Start: 145659902 End: 145854134 Strand: -	K RECOL4-007	give information about the genomic error.
Yega Havana	<pre>RECQL4-009 processed transcript < RECQL4-001 <recql4-001 <="" pre=""></recql4-001></pre>	GRC report for HG-334 There is a possible functional difference between the proteins encoded by RefSeq NM_004260.2 (RECOL4) and the genomic region to which it aligns.	processed transcript	< RECQL4-003 retained intron

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

	TTUE TOUTO
Ė٠	Other genome browsers
	- UCSC
	- NCBI
	- Vega

This will open in a new browser window.

	UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly move <<<< << >>>>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x position/search chr8:145,736,671-145,743,229 gene jump clear size 6,559 bp. configure
	chr6 (q24.3) p23.16022 8012 12.1 021.322.1 2313 024.1
Sc	cale
REC	

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	<u>tsl1</u>	
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:

Viewi 2 Genes match yo	ng 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	
ABO [Havana: O	THUMG0000020872]	as one of them is on a	
Description	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transfe Havana]	GRC patch.	essed transcript
Location	<u>9:136131053-136150617:-1</u>		
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.

Vega* BLAST/BL	AT Help & Documenta	ation					•
Human (VEGA47) ▼ Loca	tion: 9:136,131,053-136	,150,617	Gene: ABO	Transcript: ABO-00	1		
Gene-based displays				C	iene: ABO OTTHU	MG00000020872	
- Splice variants (1)	Description AB	O blood grou	p (transferase A	, alpha 1-3-N-acetylgal	actosaminyltransfe	ase; transferase B, alph	na 1-3-galactosyltransferase)
 Supporting evidence Sequence 	Location Ch	romosome 9:	136,131,053-1	36,150,617 reverse stra	nd.		
External references Comparative Genomics	Transcripts Th	is gene has 1	transcript				
 Genomic alignments Orthologues Alt. alleles 	Show/hide col	umns				Filter	
External Data	Name 🔶 Tra	nscript ID	Length (b	op) 🔶 Protein ID	🔶 Length (aa) 🍦	Biotype	CCDS 🔶
 Personal annotation Other genome browsers 	ABO-001 OTTHU	IMT00000054	1076	6 No protein prod	uct -	Processed transcript	-
- Ensembl							
Jero Configure this page					Gene summa	ry he!p	
R. Manago your data	Curated Locus	ABO (HGN	C Symbol)				
Indiage your data	Synonyms	A3GALNT,	A3GALT1 [To vi	ew all genes linked to the	ne name <u>click here</u> .	STEP 2:	
🕞 Export data	Gene type	Known proc	cessed transcrip	t [Definition]		The gene s	summary says
A Bookmark this page	Author	This gene w	vas annotated b	y Havana < <u>vega@san</u> g	er.ac.uk>	that this is	a nan aading
Dookmark this page	Version & date	Version 3, la	ast modified on	14/09/2011 (Created or	11/12/2003)	that this is	a non-cooling
	Other assemblies	This gene n Jump to this	naps to to <u>136,1</u> s stable ID in Er	31,053-136,150,617 in isembl	GRCh37 (Ensemb	transcript.	Click on the
	Curation Method	See this de	scription of the I	manual annotation proc	ess	transcript I	D to being up
	Alternative genes	Ensemble	gene: ENSG	00000175164 [view all	ocations]	the remark	s field.
			A1 -	136.12 Mb 136.13	Mb 136.13 Mb	136.14 Mb 1	36.15 Mb 136.15 Mb
	Hava	na gene		A 100 100 100 100 100 100 100 100 100 10	BO-001		

The gene is a non-coding transcript.

	Transcript: ABO-001 OTTHUMT00000054907									
Description	scription ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)									
Location	omosome 9: 136,131,053-136,150,617 reverse strand.									
Gene 🖃	is transcript is a product of gene OTTHUMG00000020872 - This gene has 1 transcript									
Show/hid	le columns Filter									
Name 🔶	Transcript ID 🖕 Length (bp) 🖕 Protein ID 🍁 Length (aa) 🖕 Biotype 🔶 CCDS 🖕									
ABO-001 0	TTHUMT00000054907 1076 No protein product - Processed transcript -									
	Transcript summary he!p									
L	Export Image									
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 symb	DOIS RP11-430N14.3-001									
Remarks	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase), ABO-*OO1 allele ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase), ABO-*OO2 allele The ABO gene in this individual produces a truncated protein without functional glycosyltransferase activity indicative of blood group O									
Other assemblie	s This transcript maps to to <u>136,131,053-136,150,617</u> in GRCh37 (Ensembl) coordinates. <u>Jump</u> to this stable ID in Ensembl									
Alternative trans	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: OT	STEP 3:	
Description	Click onto the gene ID for the HG_79 PATCH entry.	
Location	HG79_PATCH:136131200-13618	
Source	v47	

The gene is now protein coding in the patch assembly:

Gene: ABO OTTHUMG00000174691

Description Location	ABO blood group (tra Chromosome HG79	Ansferase A, alpha PATCH: 136.131	a 1-3-N-acetylgalactosami .200-136.150.736 reverse	nyltransferase; tr strand.	ansferase B, alph	a 1-3-galactosyltransfera				
Transcripts 🖃	This gene has 1 tran	is gene has 1 transcript								
Show/hide	e columns				Filter					
Name 🔶	Transcript ID 🔶	Length (bp) 🍦	Protein ID	Length (aa) 🝦	Biotype 🔶	CCDS 🔶				
ABO-001 OT	THUMT00000427219	1077	OTTHUMP00000253927	354	Protein coding	-				
Synchrynis AscAct (, AscAct (, Notact () for view all genes infect to the name circle (interact) Gene type Known protein coding [Definition] Author This gene was annotated by Havana <vega@sangerac.uk> Version & date Version 2, last modified on 17/08/2011 (Crafted on 11/08/2011) Other assemblies This gene maps to to 136,131,200-136,150,736 in GRCh37 (Ensembl) coordinates.</vega@sangerac.uk>										
Curation Method	See this descrip	tion of the manua	l annotation process							
Alternative genes	Ensembl gene	ENSG00000	256062 [view all locations]	1						
	F				9.54 Kb					
		136.12	2 Mb 136.13 Mb	136.13 Mb 1	.36.14 Mb 13	6.15 Mb 136.15 Mb				
c t	Contigs Havana gene		< ABO-001 Havana prot	AL772161. 	10.1.149482 >					

Patch assembly in Ensembl:





Warning/Error(s):

Sorry, couldn't locate ChrHG79_PATCH:136131200-136150736 in genome database



Location: HG79_F Gene:	PATCH:136131200	-136150736	Go Go				< +===			≥≫
Assembly exception Chromosome bands Primary assembly Human RefSeg/E	136,132,000	136,134,000	136,136,000	136,138,000	19.54 Kb - 136,140,000 q34.2	136,142,000 9:1360494 <u>9:1360494</u>	136,144.000 142-136369192 142-136369192	136,146,000	Forward s	136,15
Semula Ensembl/tiavana Yega.Havana. Human.RefSeq/E %GC	< ABO.1-001 protein coding < ABO-001 Vega Havana prote	in coding		-twing	STEP 6 Mouse Assemb green) t the refe Click or through reference	careful oly exce to bring rence a this to to this ce asse	ly over eption to up the assemb take yo gene in gene in embly.	the rack (irr link to ly. ou n the		

We are now back in reference human genome assembly.



	1 00 Mb Forward strand										
- A	135.80 Mb	135.90 Mb 136.00	ИЬ 136.10 МЬ	136.20 Mb	136.30 Mb	136.40 Mb	136.50 Mb	136.60 Mb 1			
Chromosome bands	q34.13				q34.2						
Contigs	AL445645.10 >	AL162417.23 >		AL158826.23 >			AL5907	10.14 >			
Ensembl/Havana											
	< AK8 ^L < TSC1 ^L C9orf9 >	GTF3C5 >	GBGT1 -< LCN1P1 -< OBP2B	L< SURF6 L	REXO4	L< TMEMBC	DBH >	<pre>'< VAV2 '< AL590710</pre>			
	GFI18	> "CEL >	L< AB(RPL7A >	LC9orf	> - FA	M163B				
	"RP11-	295G24.4.1 > "< RALGD	9	SURF1	- '< S.	C2A6	< LL09NC01-254	011.1.1			
	"RP1	1-415H23.4.1 > 4	F11-326L24.7.1 >	SURF2 5	~ RP	13-100B2.4.1	< DBH-ASJ				
		CELP >	24.4.1	< SURP	9						
		CLU P		LCN1P2 >							
				LRP11-244M	20.7.1 >						
ncRNA				'''							
	LUG > LENU	7-21P > LsnoU13 >		Y RNA >							
				- Y_RNA							
				SNORD24	4 >						
				SNORD3	6B >						
				"SNORD3	6A >						
Assembly exception			WICH BY BEE HO	SNORD3	442.3 262206	St. (End)					
STREET, STREET			PATCHLEIA REP: HO	5 PAIGH 136045	H42-1363/96	/s (mu)					

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

			Processed transcript RPL7A-008 > processed transcript	protein co	ding ADAMTS13-006 > protein coding
GRC region NCBI	HG-79 (Resolved)				
	ind to incoment				
		HG-17 (Resolved)	the second s		
Contigs	AL732364.9 >		AL158826.23 >		AL593848.14 >
Ensembl/Havana	 OBP28-001 protein coding OBP28-002 	ABO-001 processed transcript	SURF6-001 SURF4-003 protein coding processed trans SURF6-002	cript U-UI-I < REXO4-002	 SLC2A6-002 protein coding ALLO SLC2A6-003

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



Export Image

Unselected species Alpaca (Vicugna pacos) - blastz	G
Alpaca (Vicugna pacos) - blastz	0
Anole Lizard (Anolis carolinensis) - translated blat	0
Armadillo (Dasypus novemcinctus) - blastz	0
Bushbaby (Otolemur garnettii) - lastz	0
Cat (Felis catus) - blastz	0
	Armadillo (Dasypus novemcinctus) - blastz Bushbaby (Otolemur garnettii) - lastz Cat (Felis catus) - blastz

and view both genomic regions.

Location: 9:13 Gene:	6102904-136	202903		Go Go			< <	+			>>>>
					10	0.00 Kb				Forward stra	and b
Hsap Chr. 9	136	11 Mb 136	5.12 Mb 13	5.13 Mb 1	36.14 Mb 13	6.15 Mb 1	36.16 Mb	L36.17 Mb	136.18 Mb 1	36.19 Mb	136.
Assembly excepti.	-										
DCKNA.								YR	NA >		
								mis	c RNA		
Ensembl/Havana	a							1	LCN1P2 :		
									pseudoge	ene	
Contigs		AL732364.9 3					AL158826.2	23 >			
Ensembl/Havana	LICNIPI			ABO				1			GIRE
	pseudogene		D	rocessed tran	nscript			1			rotein
H.sap-H.sap lastz								1			
Assembly excepti.	PATCH FIX REP	HG79 PATCH 1	360 49 442-1 363	79605 (fwd)				1			
Hsap Chr. 9	136	11 Mb 130	512 Mb 136	5.13 Mb 1	36.14 Mb 13	6.15 Mb 1	36.16 Mb	36.17 Mb	136.18 Mb 1	36.19 Mb	136
	-Reverse st	rand —			10	0.00 Kb					
« C B TIB	2 22							1			
Conce Longood	RNA gen	0									
Gene Legend	merged	Ensembl/Hava	ana			pseudo	gene	- 1			
	- Paralogu	e						1			
					10	0.00 Kb				Forward stra	and man
Hsap HG79 PATCH	136 11 Mb	136 12 Mb	196 19 Mb	136 14 Mb	136 15 Mb	136 16 Mb	136 17 M	136 18 Mk	196 19 Mb	136.20	Mb
Assembly excepti	130.11 140	130.12 100	100.10 140	100.14 140	230.23 140	230.20 100	230.27 144	130.20 14	230.23 (10	1.50.20	MU
H sap-H sap lastz-								1			
Assembly patch n.											
CONTRACTOR AND								Y RNA >			
Ensembl Alavana								THESE RIVE	н		
	A								LCN1	L2.1 >	
Conting					AL 772	2161.10 >			pseu	dogene	
Contigs				_							
Ensempl/Havana	A.		< ABO.1								< 5
			protein co	ding					1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		pro
Assembly excepti.	PATCH_FIX: 9:1	36049442-1363	69192 (fwd)								
Hsap HG79_PATCH	136.11 Mb	136.12 Mb	136.13 Mb	136.14 Mb	136.15 Mb	136.16 Mb	136.17 M	136.18 Mb	136.19 Mb	136.20	Mb
	Heverse st	rand			10	0.00 KB					
« < 🤍 🖬 I	< > ≫										
Gene Legend	RNA gen	e									
and a state of the	protein o	oding				pseudo	gene				
	Paralogu	e									

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.