

Module 5: Working with Genome Browsers

Web-based 'genome browsers' have been developed to make it easier to access comprehensive information about regions of the human genome and about the whole human gene set. They help you to:

- Explore what is in a chromosomal region
- Search & retrieve across the whole genome
- Investigate genome organisation
- Compare to other genomes
- View alternative regions

Browsers display the location and structure of known genes and predicted novel genes along with information about the mRNA transcripts and may also include information about protein products. Information about genes is integrated with information about other genomic features (e.g. cytogenetic bands, markers, SNPs, repeated sequences, regions homologous to other species) and displayed alongside the genomic sequence assembly. Protein, mRNA and EST entries from various sequence databases may also be shown 'mapped' onto the chromosomes. Other resources that can be found include:

- **Links** to other databases and resources
- **Text Searching**
- **BLAT** and other sequence similarity searching
- **Download** of genomic sequence, gene information and other data
- **Data mining** facilities

We will take a look Biomart in Ensembl, Table Browser in the UCSC Genome Browser and biotypes and patches in Vega.

While browsers can be very useful tools, they do not provide the definitive answer to every question! Remember, new data and updates make genome browsing a fluid, changing, and improving, process.

BioMart

Demo: BioMart

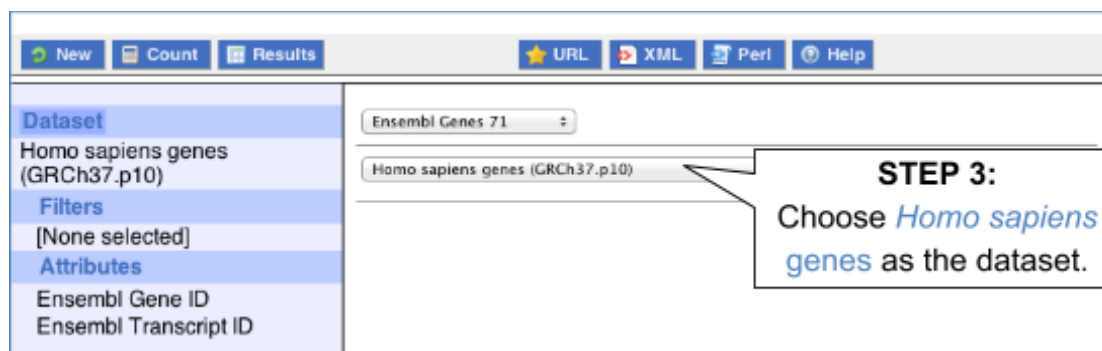
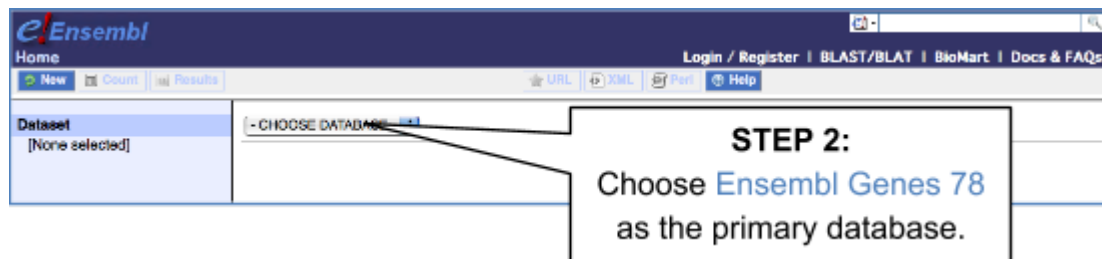
Follow these instructions to guide you through BioMart to answer the following query:

You have three questions about a set of human genes: *ESPN*, *MYH9*, *USH1C*, *CISD2*, *THRB*, *DFNB31*
(these are HGNC gene symbols. More details on the HUGO Gene Nomenclature Committee can be found on <http://www.genenames.org>)

- 1) What are the EntrezGene IDs for these genes?
- 2) Are there associated functions from the GO (gene ontology) project that might help describe their function?
- 3) What are their cDNA sequences?

Step 1: Click on [BioMart](#) in the top header of a www.ensembl.org page to go to: www.ensembl.org/biomart/martview

NOTE: These answers were determined using BioMart Ensembl 79.



STEP 4:
Click **Filters** at the left.
Expand the **GENE** panel.

The screenshot shows the BioMart interface with the 'Filters' panel on the left containing 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The main query area has the 'GENE' filter expanded, showing options like 'Limit to genes (external references)...' and 'Input external references ID list [Max 500 advised]'.

STEP 5:
In **Input external references ID list**,
paste in your gene symbols. Change
the heading to read **Associated Gene
Name(s) [e.g. BRCA2]**

The screenshot shows the 'Input external references ID list' field with a dropdown menu containing gene symbols: BRCA2, TP53, MYH4, USH1G, GSD2, THRB, GFM31. The heading is 'Associated Gene Name(s) [e.g. BRCA2]'.

STEP 6:
Click **Count** to see BioMart is reading 6
genes out of 65,803 possible *Homo
sapiens* genes. Since we entered 6
gene symbols, this confirms that our
filters have worked correctly.

The screenshot shows the 'Count' button highlighted in the top navigation bar. The 'Dataset' panel displays '6 / 63292 Genes' and 'Homo sapiens genes (GRCh38)'. The 'Filters' panel shows 'HGNC symbol(s) [e.g. NTN3]: [ID-list specified]'.

STEP 7:
Click on **Attributes** to
select output options
(i.e. GO terms)

STEP 8:
Expand the **EXTERNAL**
and **GENE** panels.

The screenshot shows the 'Attributes' panel selected in the left sidebar, with options for 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The main query area has the 'EXTERNAL' and 'GENE' filters expanded. The 'EXTERNAL' filter has radio buttons for 'Features', 'Structures', 'Transcript Event', 'Homologs', 'Variation', and 'Sequences'. The 'GENE' filter is also expanded.

LRG to Ensembl link gene
 LRG to Ensembl link transcript
 EntrezGene ID
 EntrezGene transcript name
 Human Protein Atlas Antibody ID

STEP 10:
 Scroll down to select
EntrezGene ID
 (to answer question 1)

GENE:

EXTERNAL:

GO

 GO Term Accession
 GO Term Name
 GO Term Definition

STEP 11:
 Scroll back up to
 select **GO term fields**
 (to answer question 2)

STEP 12:
 Click **Results**.

Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	HGNC symbol	GO Term Accession	GO Term Name	GO Term Definition
ENSG00000151090	ENST00000356447	7065	THPR	GO:000122	negative regulation of transcription from RNA polymerase II promoter	"Any process that stops, prevents, or reduces the frequency, rate or extent of transcription from an RNA polymerase II promoter." [GOC:curators, GOC:txnh]
ENSG00000151090	ENST00000356447	7065	THPR	GO:000351	transcription, DNA-dependent	"The cellular synthesis of RNA on a template of DNA." [GOC:curators, GOC:txnh]
ENSG00000151090	ENST00000356447	7065	THPR	GO:004544	positive regulation of transcription from RNA polymerase II promoter	"Any process that activates or increases the frequency, rate or extent of transcription from an RNA polymerase II promoter." [GOC:curators, GOC:txnh]

Why are there multiple rows for one gene ID? For example, look at the first few rows.

Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	GO Term Accession	GO Term Name	GO Term Definition	HGNC symbol
ENSG00000187017	ENST00000377828	83715	GO:0007605	sensory perception of sound	"The series of events required for an organism to receive an auditory stimulus, convert it to a molecular signal, and recognize and characterize the signal. Sonic stimuli are detected in the form of vibrations and are processed to form a sound." [GOC:ai]	ESPN
ENSG00000187017	ENST00000377828	83715	GO:0007626	locomotory behavior	"The specific movement from place to place of an organism in response to external or internal stimuli. Locomotion of a whole organism in a manner dependent upon some combination of that organism's internal state and external conditions." [GOC:dph]	ESPN
ENSG00000187017	ENST00000377828	83715	GO:0030046	parallel actin filament bundle assembly	"Assembly of actin filament bundles in which the filaments are tightly packed (approximately 10-20 nm apart) and oriented with the same polarity." [GOC:mah, ISBN:0815316194]	ESPN

STEP 13:
Click **Attributes** again

STEP 14:
Select **Sequences** at the top, then expand **SEQUENCES** and choose the option **cDNA sequences** (to answer question 3).

STEP 15:
Expand **Header Information** to select the **Associated Gene Name**.

STEP 16:
Click **Results** to see the cDNA sequences in FASTA format.

STEP 17:
Change **View 10 rows** to **View All rows** so that you see the full table.
Note: Pop-up blocking must be switched off in your browser.

What did you learn about the human genes in this exercise?
 Could you learn these things from the Ensembl browser? Would it take longer?

For more details on BioMart, have a look at these publications:

Smedley, D. *et al* **BioMart – biological queries made easy**

BMC Genomics 2009 Jan 14;10:22

Kinsella, R.J. *et al* **Ensembl BioMart: a hub for data retrieval across taxonomic space.**








Database (Oxford) 2011:bar03

Uploading data to Ensembl

Demo: Attach URLs of large files

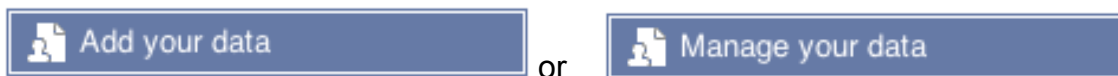
Large files, such as BAM files generated by NGS, need to be attached by URL to be viewed in Ensembl. I've put a BAM file of human chromosome 20 RNASeq data online at: <http://www.ebi.ac.uk/~emily/Workshops/BAM/>

Let's take a look at that URL.

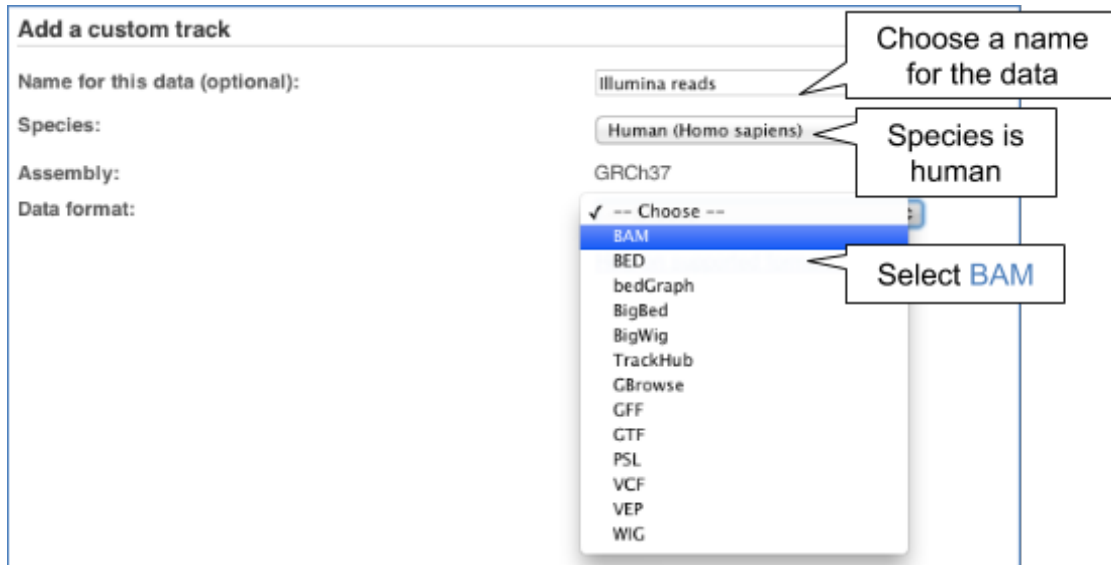
Index of /~emily/Workshops/BAM			
Name	Last modified	Size	Description
 Parent Directory		-	
 GRCh38.20.illumina.merged.1.bam	25-Jul-2014 15:08	2.8G	
 GRCh38.20.illumina.merged.1.bam.bai	25-Jul-2014 15:08	169K	
 GRCh38.21.illumina.merged.1.bam	25-Jul-2014 15:17	2.9G	
 GRCh38.21.illumina.merged.1.bam.bai	25-Jul-2014 15:17	121K	
 Illumina reads test.bam	19-Apr-2013 14:17	394M	
 Illumina reads test.bam.bai	19-Apr-2013 14:16	176K	

Here you can see a number of BAM files (.bam) with corresponding index files (.bam.bai). We're interested in the files [GRCh38.20.illumina.merged.1.bam](#) and [GRCh38.20.illumina.merged.1.bam.bai](#). These files are the BAM file and the index file respectively. When attaching a BAM file to Ensembl, there must be an index file in the same folder.

Click on the [Add your data](#) button at the left. If you've previously added data to Ensembl, this button will say [Manage your data](#) instead.

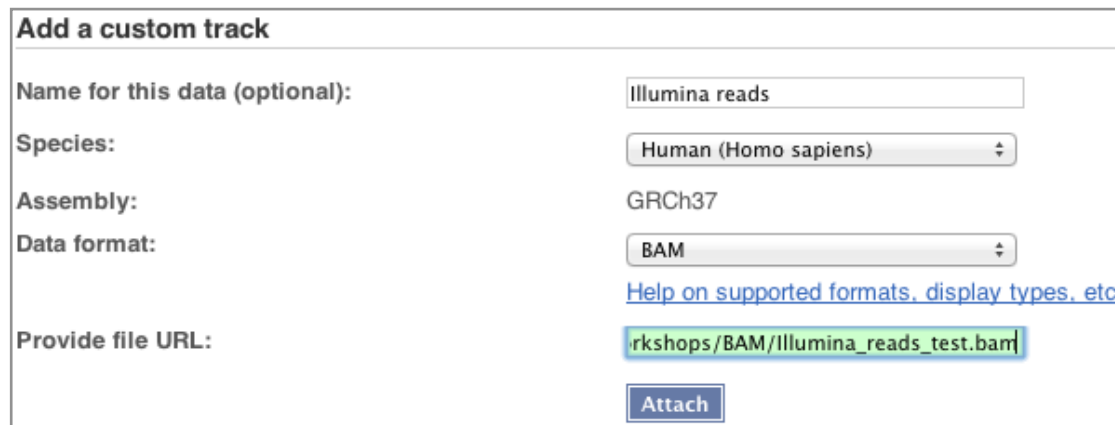


A menu will appear:



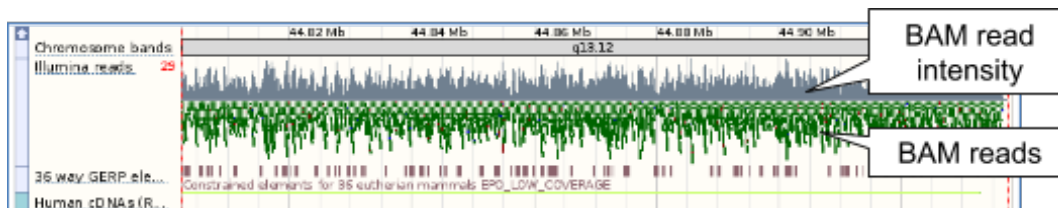
We'll name our data **Illumina reads** and choose **BAM** as the data format.

Paste in the URL of the BAM file itself (<http://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam>), then click **Attach**.

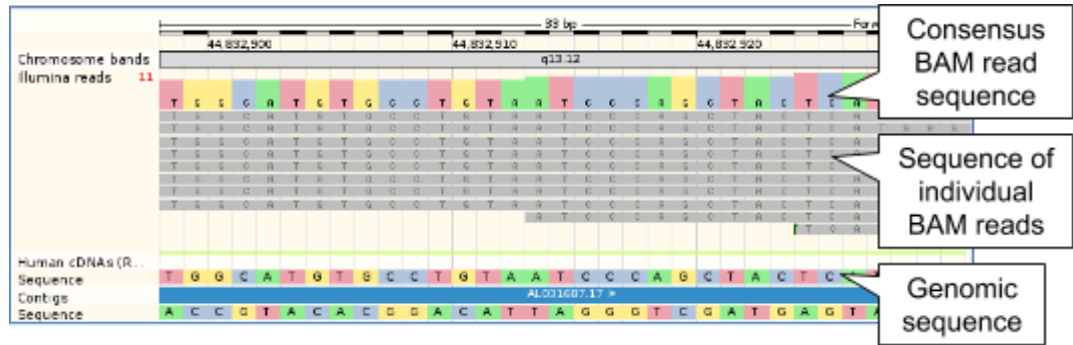


Close the menu.

To see this data, jump to a region on chromosome 20. Let's go to the region of the *CDH22* gene. Search for the gene and click on the location.



We can zoom in to see the sequence itself. Drag out boxes in the view to zoom in, until you see a view like this.



Exercises: BioMart**Exercise 1 – BioMart: Finding genes by protein domain**

Find mouse proteins with transmembrane domains located on chromosome 9.

Exercise 2 – BioMart: Convert IDs

BioMart is a very handy tool when you want to convert IDs from different databases. The following is a list of 29 IDs of **human proteins** from the NCBI **RefSeq** database (<http://www.ncbi.nlm.nih.gov/projects/RefSeq/>):

NP_001218	NP_203125
NP_203124	NP_203126
NP_001007233	NP_150636
NP_150635	NP_001214
NP_150637	NP_150634
NP_150649	NP_001216
NP_116787	NP_001217
NP_127463	NP_001220
NP_004338	NP_004337
NP_116786	NP_036246
NP_116756	NP_116759
NP_001221	NP_203519
NP_001073594	NP_001219
NP_001073593	NP_203520
NP_203522	

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq IDs correspond. Do these 29 proteins correspond to 29 genes?

Hint: For this exercise, it's easier to copy and paste the IDs from the online exercise booklet (copy one column, then the other).

Exercise 3 – BioMart: Export homologues

For a list of *Ciona savignyi* Ensembl genes, export the human orthologues.

ENSCSAVG000000000002
 ENSCSAVG000000000003
 ENSCSAVG000000000006
 ENSCSAVG000000000007
 ENSCSAVG000000000009

ENSCSAVG00000000011

Exercise 4 – BioMart: Find genes associated with array probes

Forrest *et al* performed a microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers (Environ Health Perspect. 2005 June; 113(6): 801–807). The microarray used was the human Affymetrix U133A/B (also called U133 plus 2) GeneChip. The top 25 up-regulated probe-sets were:

207630_s_at	221840_at
219228_at	204924_at
227613_at	223454_at
228962_at	214696_at
210732_s_at	212370_at
225390_s_at	227645_at
226652_at	221641_s_at
202055_at	226743_at
228393_s_at	225120_at
218515_at	202224_at
200614_at	212014_x_at
223461_at	209835_x_at
213315_x_at	

- (a) Retrieve for the genes corresponding to these probe-sets the Ensembl Gene and Transcript IDs as well as their HGNC symbols and descriptions.
- (b) In order to analyse these genes for possible promoter/enhancer elements, retrieve the 2000 bp upstream of the transcripts of these genes.
- (c) In order to be able to study these human genes in mouse, identify their mouse orthologues. Also retrieve the genomic coordinates of these orthologues.

Exercise 5 – BioMart: Export structural variants

You can use BioMart to query variants, not just genes. (Make sure you use the right Datasets.)

- (a) Export the study accession, source name, chromosome, sequence region start and end (in bp) of human structural variations (SV) on chromosome 1, starting at 130,408 and ending at 210,597.
- (b) In a new BioMart query, find the alleles, phenotype descriptions, and associated genes for rs1801500 and rs1801368. Can you view this same information in the Ensembl browser?

Exercise Answers:

Exercise 1 – BioMart: Finding genes by protein domain

As with all BioMart queries you must select the [dataset](#), set your [filters](#) (input) and define your [attributes](#) (desired output). For this exercise:

Dataset: Ensembl genes in mouse

Filters: Transmembrane proteins on chromosome 9

Attributes: Ensembl gene and transcript IDs and Associated gene names

Go to the Ensembl homepage (<http://www.ensembl.org>) and click on [BioMart](#) at the top of the page.

Select [Ensembl genes](#) as your database and [Mus musculus genes](#) as the dataset.

Click on [Filters](#) on the left of the screen and expand [REGION](#). Change the [chromosome](#) to [9](#).

Now expand [PROTEIN DOMAINS](#), also under filters, and select [Limit to genes](#), choosing [with Transmembrane domains](#) from the drop-down and then [Only](#). Clicking on [Count](#) should reveal that you have filtered the dataset down to 425 genes.

Click on [Attributes](#) and expand [GENE](#). Select [Associated gene name](#).

Now click on [Results](#). The first 10 results are displayed by default; display all results by selecting [ALL](#) from the drop down menu.

The output will display the Ensembl gene ID, Ensembl Transcript ID and Associated gene names of all proteins with a transmembrane domain on mouse chromosome 9. If you prefer, you can also export as an Excel sheet by using the Export all results to XLS option.

Exercise 2 – BioMart: Convert IDs

Click [New](#).

Choose the [ENSEMBL Genes 79](#) database.

Choose the [Homo sapiens genes \(GRCh38\)](#) dataset.

Click on [Filters](#) in the left panel.

Expand the [GENE](#) section by clicking on the [+](#) box.

Select [Input external references ID list - RefSeq protein ID\(s\)](#) and enter the list of IDs in the text box (either comma separated or as a list).

HINT: You may have to scroll down the menu to see these.

[Count](#) shows 11 genes (remember one gene may have multiple splice variants coding for different proteins, that is the reason why these 29 proteins do not correspond to 29 genes).

Click on [Attributes](#) in the left panel.
Select the [Features](#) attributes page.
Expand the [External section](#) by clicking on the + box.
Select [HGNC symbol](#) and [RefSeq Protein ID](#) from the [External References](#) section.

Click the [Results](#) button on the toolbar.
Select [View All rows as HTML](#) or export all results to a file.

Exercise 3 – BioMart: Export homologues

Click [New](#).
Choose the [ENSEMBL Genes 79](#) database.
Choose the [Ciona savignyi genes \(CSAV2.0\)](#) dataset.

Click on [Filters](#) in the left panel.
Expand the [GENE](#) section by clicking on the + box.
Enter the gene list in the [Input external references ID list](#) box.

Click on [Attributes](#) in the left panel.
Select the [Homologs](#) attributes page.
Expand the [Orthologs](#) section by clicking on the + box.
Select [Human Ensembl Gene ID](#).
Click [Results](#).

Exercise 4 – BioMart: Find genes associated with array probes

(a) Click [New](#).
Choose the [ENSEMBL Genes 79](#) database.
Choose the [Homo sapiens genes \(GRCh38\)](#) dataset.

Click on [Filters](#) in the left panel.
Expand the [GENE](#) section by clicking on the + box.
Select [Input microarray probes/probesets ID list - Affy hg u133 plus 2 probeset ID\(s\)](#) and enter the list of probeset IDs in the text box (either comma separated or as a list).

[Count](#) shows 24 genes match this list of probesets.

Click on [Attributes](#) in the left panel.
Select the [Features](#) attributes page.
Expand the [GENE](#) section by clicking on the + box.
In addition to the default selected attributes, select [Description](#).
Expand the [External](#) section by clicking on the + box.

Select [HGNC symbol](#) from the [External References](#) section and [AFFY HG U133-PLUS-2](#) from the [Microarray Attributes](#) section.

Click the [Results](#) button on the toolbar.

Select [View All rows as HTML](#) or export all results to a file. Tick the box [Unique results only](#).
Your results should show that the 25 probes map to 24 Ensembl genes.

(b) Don't change Dataset and Filters – simply click on [Attributes](#).

Select the [Sequences](#) attributes page.

Expand the [SEQUENCES](#) section by clicking on the + box.

Select [Flank \(Transcript\)](#) and enter [2000](#) in the [Upstream flank](#) text box.

Expand the [Header information](#) section by clicking on the + box.

Select, in addition to the default selected attributes, [Description](#) and [Associated Gene Name](#).

Note: Flank (Transcript) will give the flanks for all transcripts of a gene with multiple transcripts. Flank (Gene) will give the flanks for one possible transcript in a gene (the most 5' coordinates for upstream flanking).

Click the [Results](#) button on the toolbar.

(c) You can leave the Dataset and Filters the same, and go directly to the Attributes section:

Click on [Attributes](#) in the left panel.

Select the [Homologs](#) attributes page.

Expand the [GENE](#) section by clicking on the + box.

Select [Associated Gene Name](#).

Deselect [Ensembl Transcript ID](#).

Expand the [ORTHOLOGS](#) section by clicking on the + box.

Select [Mouse Ensembl Gene ID](#), [Mouse Chromosome Name](#), [Mouse Chr Start \(bp\)](#) and [Mouse Chr End \(bp\)](#).

Click the [Results](#) button on the toolbar.

Select [View All rows as HTML](#) or export all results to a file.

Your results should show that for most of the human genes at least one mouse orthologue has been identified.

Exercise 5 – BioMart: Export structural variants

(a) Choose [Ensembl Variation 79](#) and [Homo sapiens Structural Variation \(GRCh38\)](#).

Filters: Region: [Chromosome 1](#), Base pair start: [130408](#), Base pair end: [210597](#)

Count shows 35 out of 4,163,079 structural variants.

Attributes: [Structural Variation \(SV\) Information: DGVa Study Accession](#) and [Source Name](#)

Structural Variation (SV) Location: Chromosome name, Sequence region start (bp) and Sequence region end (bp).

(b) Choose Ensembl Variation 79 and Homo sapiens Short Variation (SNPs and indels) (GRCh38).

Filters: Filter by Variation name enter: **rs1801500**, **rs1801368**

Attributes: Variation Name, Variant Alleles, Phenotype description and Associated gene.

You can view this same information in the Ensembl browser. Click on one of the variation IDs (names) in the result table. The variation tab should open in the Ensembl browser. Click [Phenotype Data](#).

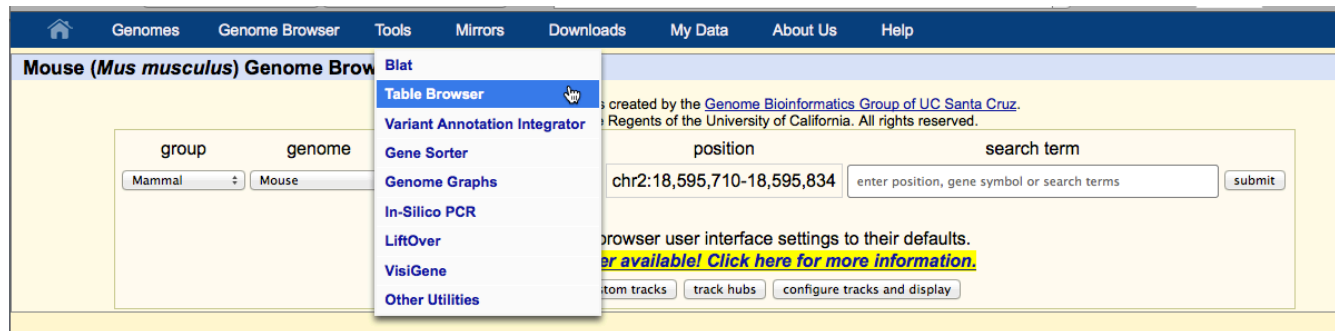
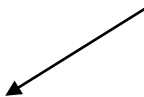
The UCSC Table Browser:

The underlying data for the UCSC browser is arranged in primary (positions, names etc) and auxiliary tables within a MySQL database. This data can be queried using table browser.

Worked example 1:

In this example we'll find the number of UCSC genes in the ENCODE pilot regions on the human genome with more than 20 exons.

STEP 1:
Click on Tools and then Table Browser



Select corresponding database table for track. In this case knownGene

Select annotation track to extract data from. For this example use UCSC Genes

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help and controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features on a [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and their entirety from the [Sequence and Annotation Downloads](#) page.

clade: genome: assembly:

group: track:

table:

region: genome ENCODE Pilot regions position

identifiers (names/accessions):

filter:

intersection with knownGene:

correlation:

output format: Send output to [Galaxy](#)

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

Note: The all fields and selected fields output formats are not available when an intersection has been specified.

To reset all user can settings (including custom tracks), [click here](#).

Select specific region, ENCODE regions, or whole genome. Select Encode pilot regions

Can upload list of name/accessions to retrieve data for

Can filter based on values in certain fields

Can output in many different ways, including getting sequence

STEP 2:
Click on "create" button next to filter

String fields that can use wildcards '*'

Numeric fields with range of =, <, >, etc

Free form query field is for SQL queries

Filter on Fields from hg19.knownGene

name	does	match	*	
chrom	does	match	*	AND
strand	does	match	*	AND
txStart	is	ignored	0	AND
txEnd	is	ignored	0	AND
cdsStart	is	ignored	0	AND
cdsEnd	is	ignored	0	AND
exonCount	is	>	20	AND
exonStarts	does	match	*	
exonEnds	does	match	*	
proteinID	does	match	*	AND
alignID	does	match	*	AND

AND Free-form query:

STEP 3:
Click on exonCount is and change to > then add in 20. Click on submit.

Click on [summary/statistics](#) to give the number of genes found:

UCSC Genes (knownGene) Summary Statistics

item count	159
item bases	4,365,689 (14.57%)
item total	20,810,028 (69.47%)
smallest item	12,124
average item	130,881
biggest item	1,806,764
block count	4,213
block bases	298,088 (1.00%)
block total	839,651 (2.80%)
smallest block	6
average block	199
biggest block	11,938

Region and Timing Statistics

region	encode
bases in region	29,955,196
bases in gaps	0
load time	4.33
calculation time	0.01
free memory time	0.00
filter	on
intersection	off

STEP 4:
Click back to table browser and under output format choose selected fields from primary and related tables. Then click get output. Leave output file blank.

filter:

intersection:

correlation:

output format: Send output to [Galaxy](#) [GREAT](#)
 all fields from selected table
 selected fields from primary and related tables
 sequence
 GTF - gene transfer format
 CDS FASTA alignment from multiple alignment
 BED - browser extensible data
 custom track
 hyperlinks to Genome Browser

output file:

file type return:

To reset all user cart settings (including custom tracks), [click here](#).

STEP 5:
Select the fields shown. Then click get output.

Select Fields from hg19.knownGene

<input checked="" type="checkbox"/>	name	Name of gene
<input checked="" type="checkbox"/>	chrom	Reference sequence chromosome or scaffold
<input type="checkbox"/>	strand	+ or - for strand
<input checked="" type="checkbox"/>	txStart	Transcription start position
<input checked="" type="checkbox"/>	txEnd	Transcription end position
<input type="checkbox"/>	cdsStart	Coding region start
<input type="checkbox"/>	cdsEnd	Coding region end
<input type="checkbox"/>	exonCount	Number of exons
<input type="checkbox"/>	exonStarts	Exon start positions
<input type="checkbox"/>	exonEnds	Exon end positions
<input checked="" type="checkbox"/>	proteinID	UniProt display ID for Known Genes, UniProt accession or RefSeq protein ID for UCSC Genes
<input type="checkbox"/>	alignID	Unique identifier for each (known gene, alignment position) pair

hg19.kgXref fields

<input type="checkbox"/>	kgID	Known Gene ID
<input type="checkbox"/>	mRNA	mRNA ID
<input type="checkbox"/>	spID	UniProt protein Accession number
<input type="checkbox"/>	spDisplayID	UniProt display ID
<input checked="" type="checkbox"/>	geneSymbol	Gene Symbol
<input type="checkbox"/>	refseq	RefSeq ID
<input type="checkbox"/>	protAcc	NCBI protein Accession number


```
#filter: knownGene.exonCount > 20
#hg19.knownGene.name      hg19.knownGene.chrom      hg19.knownGene.txStart    hg19.kno
uc003vij.3                chr7                       116312458                 116438440                 P08581 MET
uc010lkh.3                chr7                       116312458                 116438440                 P08581-2 MET
uc011knj.2                chr7                       116364175                 116438440                 P08581 MET
uc003vjd.3                chr7                       117120016                 117308718                 P13569 CFTR
uc011knq.2                chr7                       117120016                 117308718                 P13569 CFTR
uc003vjf.3                chr7                       117350705                 117513561                 Q8WZ74 CTTNBP2
uc003kvv.1                chr5                       131142839                 131329971                 ACSL6
uc010jdn.2                chr5                       131285666                 131329944                 Q9UKU0-6 ACSL6
uc003kwb.3                chr5                       131285666                 131347355                 NP_001192176 ACSL6
uc003kvx.2                chr5                       131285666                 131347355                 Q9UKU0-8 ACSL6
uc003kvy.2                chr5                       131285666                 131347355                 Q9UKU0-1 ACSL6
uc010jdo.2                chr5                       131285666                 131347607                 Q9UKU0-3 ACSL6
uc003kwa.2                chr5                       131285666                 131347761                 NP_001192179 ACSL6
uc003kxi.3                chr5                       131892615                 131980313                 Q92878 RAD50
uc003kxh.3                chr5                       131892615                 131980313                 Q92878 RAD50
uc003kyd.3                chr5                       132211070                 132299354                 Q9UHB7 AFF4
uc011cxk.2                chr5                       132211070                 132299354                 Q9UHB7 AFF4
uc001ppy.3                chr11                     116714117                 116968993                 Q9Y2K2 SIK3
uc001ppz.3                chr11                     116714117                 116968993                 A1A5A9 SIK3
uc001pqa.3                chr11                     116714117                 116968993                 A1A5A8 SIK3
uc003ale.3                chr22                     31892260                  32014534                  A8K8P3 SFI1
uc003alf.3                chr22                     31892260                  32014534                  A8K8P3-2 SFI1
uc003alg.3                chr22                     31892260                  32014534                  A8K8P3-3 SFI1
uc011alp.2                chr22                     31892260                  32014534                  A8K8P3-10 SFI1
uc011alq.2                chr22                     31892260                  32014534                  A8K8P3-9 SFI1
```

Output sent to browser window.

Worked example 2:

Search for the number of simple repeats on human chromosome 4 between 3 and 4 million bp that have a copy number of more than 10. Then find out how many of these simple repeats are located in known genes.

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence coverage. [Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. Through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page [Sequence and Annotation Downloads](#) page.

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)

group: Repeats track: Simple Repeats manage custom tracks track hubs

table: simpleRepeat describe table schema

region: genome ENCODE Pilot regions position chr4:3000000-4000000 lookup define regions

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection with knownGene: edit clear

correlation: create

output format: custom track Send output to Galaxy GREAT GenomeSpace

output file: example.bed (leave blank to keep output in browser)

file type returned: plain text gzip compressed

Note: The all fields and selected fields output formats are not available when an intersection has been specified.

get output summary/statistics

To reset all user cart settings (including custom tracks), [click here](#).

STEP1:
 Select:
 group – Variation and Repeats
 track – Simple Repeats
 table - simpleRepeat
 region – position 3000000-4000000

Click on create filter.

bin	is	ignored	0	
chrom	does	match	*	AND
chromStart	is	ignored	0	AND
chromEnd	is	ignored	0	AND
name	does	match	*	
period	is	ignored	0	
copyNum	is	>	10	AND
consensusSize	is	ignored	0	AND
perMatch	is	ignored	0	AND
perIndel	is	ignored	0	AND
score	is	ignored	0	AND
A	is	ignored	0	AND
C	is	ignored	0	AND
G	is	ignored	0	AND
T	is	ignored	0	AND
entropy	is	ignored	0	AND
sequence	does	match	*	

AND Free-form query:

STEP 2:
Select copyNum is > 10.
Then click submit

STEP 3:
Click on summary/statistics to get the result of 140. Then go back in your browser to get back to table browser.

Simple Repeats (simpleRepeat) Summary

item count	140
item bases	31,262 (3.13%)
item total	54,691 (5.47%)
smallest item	25
average item	391
biggest item	1,869
smallest score	50
average score	290
biggest score	2,542

Region and Timing Statistics

region	chr4:3000000-4000000
bases in region	1,000,001
bases in gaps	0
load time	0.07
calculation time	0.00
free memory time	0.00
filter	on
intersection	off

STEP 4:
Click on create intersection

intersection:

This will create an intersection with another data set, that is anything that overlaps. For this query we will select known genes.

Intersect with Simple Repeats

Select a group, track and table to intersect with:

group: track:
 table:

Note: UCSC Genes has gene/alignment structure. Only the bases covered by its exons/blocks will be considered.

Intersect Simple Repeats items with bases covered by UCSC Genes:

These combinations will maintain the names and gene/alignment structure (if any) of Simple Repeats:

- All Simple Repeats records that have any overlap with UCSC Genes
- All Simple Repeats records that have no overlap with UCSC Genes
- All Simple Repeats records that have at least % overlap with UCSC Genes
- All Simple Repeats records that have at most % overlap with UCSC Genes

Intersect bases covered by Simple Repeats and/or UCSC Genes:

These combinations will discard the names and gene/alignment structure (if any) of Simple Repeats and produce a simple list of position ranges.

- Base-pair-wise intersection (AND) of Simple Repeats and UCSC Genes
- Base-pair-wise union (OR) of Simple Repeats and UCSC Genes

Check the following boxes to complement one or both tables. To complement a table means to include a base pair in the intersection/union if it is *not* included in the table.

- Complement Simple Repeats before base-pair-wise intersection/union
- Complement UCSC Genes before base-pair-wise intersection/union

STEP 5:
Click on submit, then get summary/statistics.

Simple Repeats (simpleRepeat) Summary Statistics

item count	5
item bases	2,238 (0.22%)
item total	3,612 (0.36%)
smallest item	64
average item	722
biggest item	1,380
smallest score	61
average score	954
biggest score	1,712

There are 5 known UCSC genes that contain simple repeats that have a copy number of > 10.

Region and Timing Statistics

region	chr4:3000000-4000000
bases in region	1,000,001
bases in gaps	0
load time	0.04
calculation time	0.00
free memory time	0.00
filter	on
intersection	on

STEP 6:
Click back to table browser then select custom track in output format and give the track a name ending in .bed.
Click on get output

This can be viewed as a custom track in UCSC.

output format: Send output to Galaxy GREAT

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

Note: The all fields and selected fields output formats are not available when an intersection has been specified.

Output simpleRepeat as Custom Track

Custom track header:

name=

description=

visibility=

url=

Create one BED record per:

Whole Gene

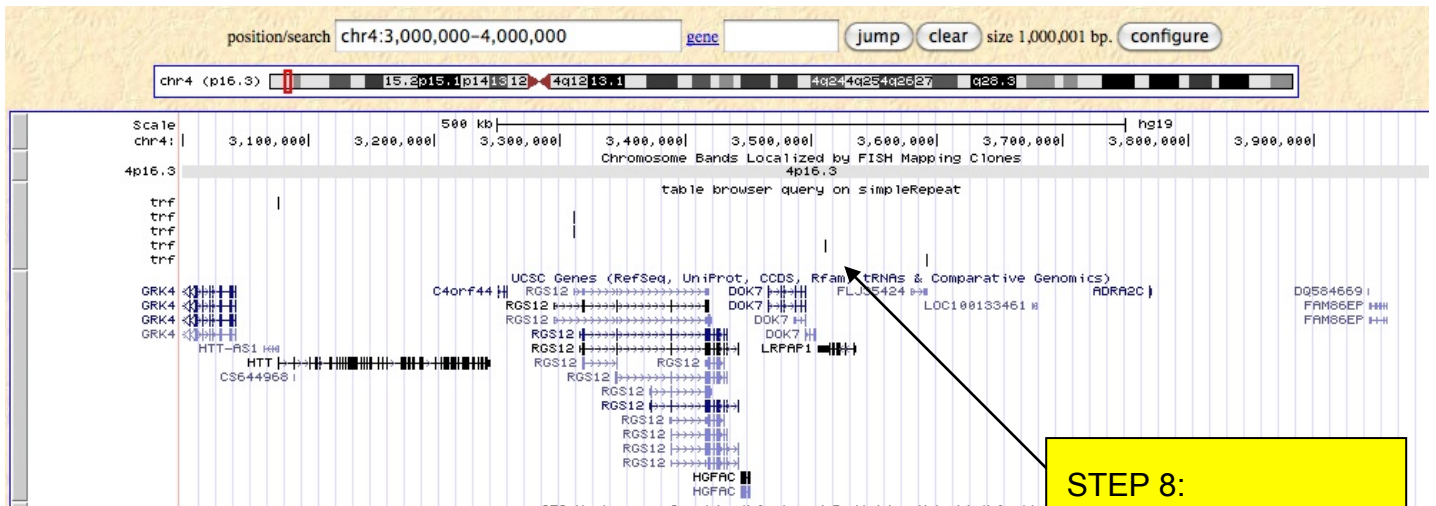
Upstream by bases

Downstream by bases

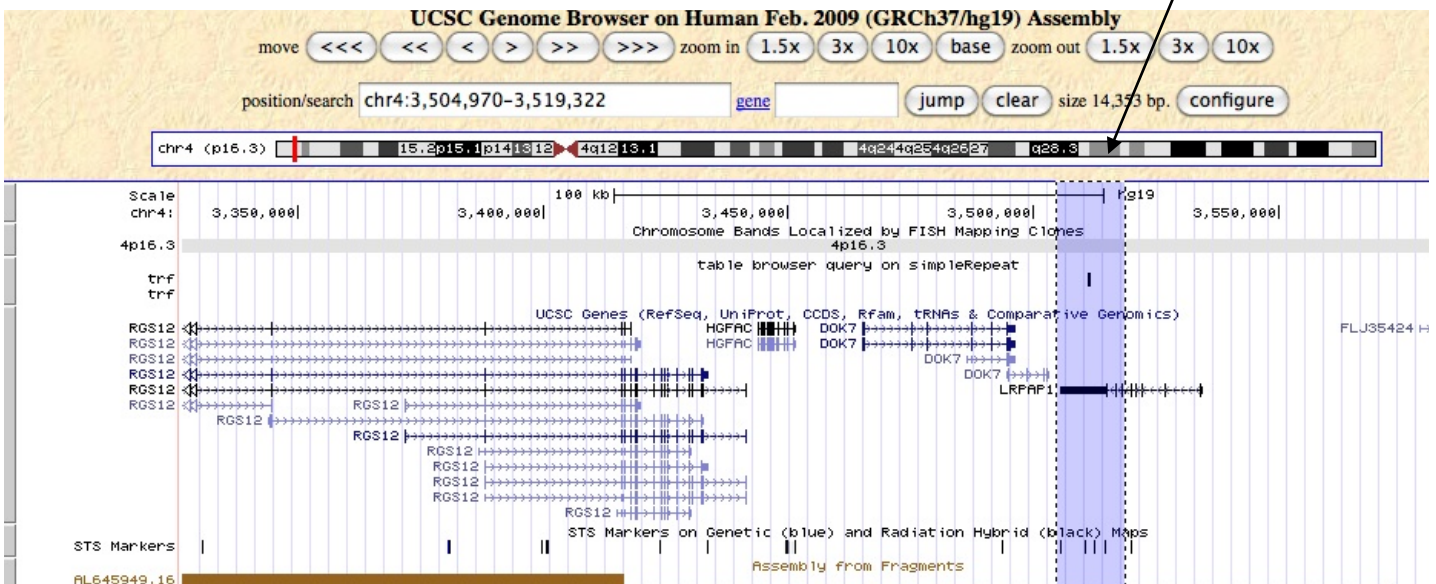
STEP 7:
Click get custom track in genome browser

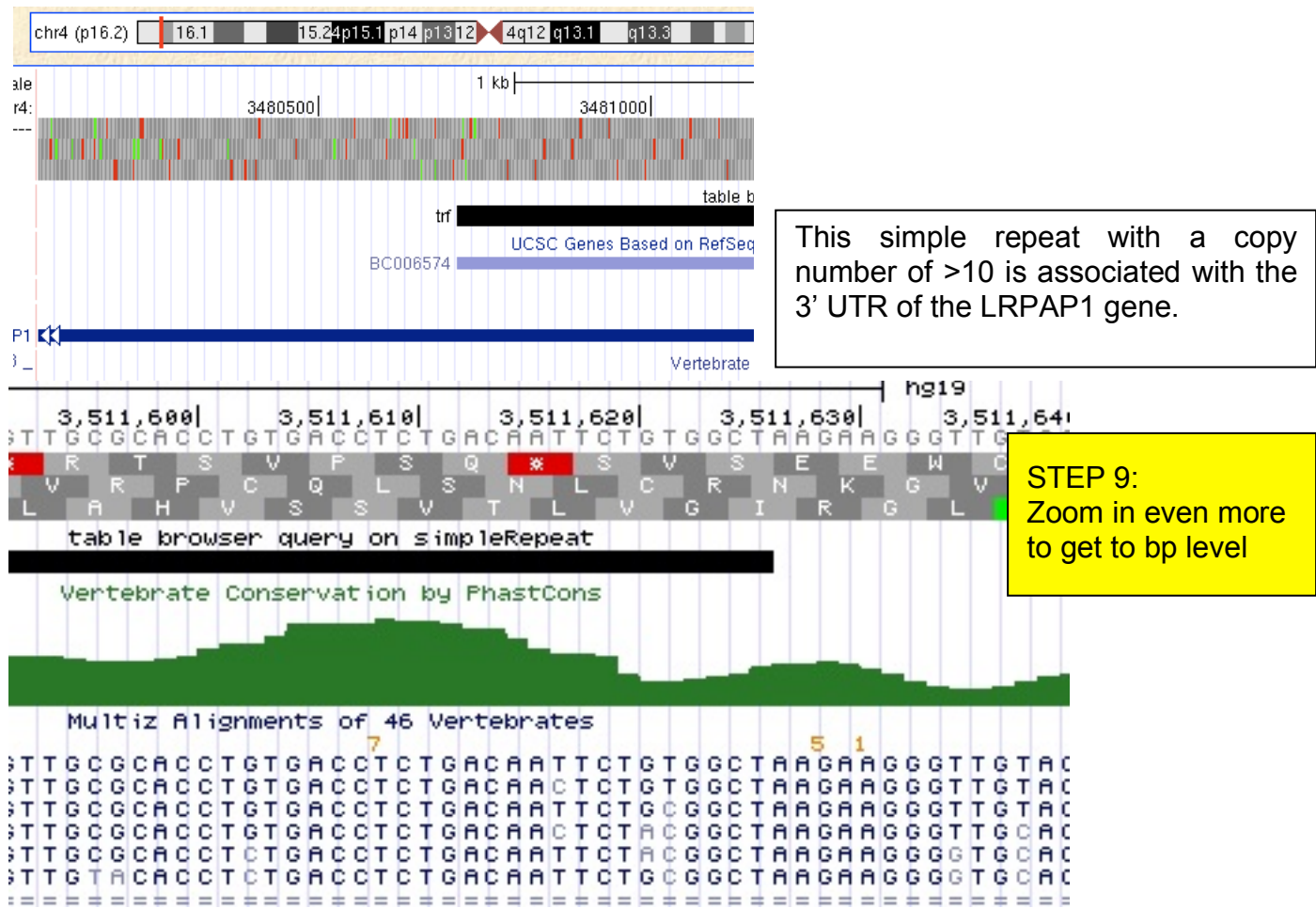
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are

You can make a file of the custom track for later use, as custom tracks are only available for 8 hours on the browser.
Get custom track in table browser will mean that you can use the track for another intersection.



STEP 8:
Zoom in on any of the 5 regions in the track by drawing a box around them.





There are many other functions in table browser. Examples include: correlation, to calculate a simple linear regression between two datasets. You may also use your own data in a track (see Custom Tracks on the browser for details). You may also load your own GWAS data (see Genome Graphs), but there are also many custom tracks from outside data sources already available (see Custom tracks) e.g. GWAS of bipolar disorder, CNVs, structural RNAs etc.

Tasks:

1. Obtain a list of SNPs in a single gene (CIZ1) using table browser
2. Find all the genes on human chromosome 22, add the gene symbols and GO IDs using table browser.

Answers:

1.

Go to table browser, select human and February 2009 assembly. Choose group variation and repeats in group, and All SNPs135 in the track menu and snp135 in the table menu.

Type in CIZ1 in the position box and then click lookup. The second entry (uc011mar.2) gives the longest transcript with the position of chr9:130, 928, 344-130, 953, 868. Under output format choose selected fields from primary and related tables, then click get output. In the menu then choose chrom, chromstart, chromend, name, strand, observed and func. Click get output.

#chrom	chromStart	chromEnd	name	strand	observed	func
chr9	130928357	130928358	rs12376026	+	A/G	untranslated-3
chr9	130928426	130928427	rs186647759	+	C/T	untranslated-3
chr9	130928567	130928568	rs192626276	+	C/G	missense
chr9	130928610	130928611	rs11549264	-	C/T	coding-synon
chr9	130928632	130928633	rs11549260	+	C/T	missense
chr9	130928652	130928653	rs184638839	+	A/G	coding-synon
chr9	130928688	130929040	rs71705963	+	(LARGEDELETION)/-	frameshift
chr9	130928721	130928722	rs188910316	+	G/T	intron
chr9	130928811	130928812	rs45437098	-	A/G	intron
chr9	130928826	130928827	rs45542735	-	C/T	intron
chr9	130928881	130928882	rs141569452	+	C/T	intron
chr9	130928972	130928973	rs45487100	-	A/G	intron
chr9	130929029	130929030	rs41276232	+	C/G/T	intron
chr9	130929138	130929139	rs140324491	+	C/T	missense
chr9	130929139	130929140	rs185085914	+	A/G	coding-synon
chr9	130929155	130929156	rs15126	+	A/T	missense

2.

Go to table browser, select human and February 2009 assembly. Choose genes and gene prediction tracks group and the track UCSC genes. Then select the table knownGene. Choose the position button and type chr22 then click lookup. This adds the range for the whole chromosome. The output needs to be selected fields from primary and related tables. Select name, chrom and protein ID. Then add some fields from the hg19.kgXref fields box, namely kgID, geneSymbol and refseq. Selecting fields in the kgXref table has now made new tables available in the linked tables area below. Check the go section, which is at the top of the linked tables, and is called goaPart. Then click on “allow selection from checked tables” at the bottom of the page. Select gold to get all the GO IDs, then select get output from the section above the go.goaPart fields table.

#hg19.knownGene.name	hg19.knownGene.chrom	hg19.knownGene.proteinID	go.goaPart.goId	hg19.kgXref.kgID	hg19.kgXref.geneS
uc002zks.4	chr22	n/a	uc002zks.4	AK022914	
uc002zkt.3	chr22	n/a	uc002zkt.3	BC040855	
uc002zku.3	chr22	n/a	uc002zku.3	BC017398	
uc002zkv.3	chr22	n/a	uc002zkv.3	AK056135	
uc021wkd.1	chr22	n/a	uc021wkd.1	DQ590589	
uc002zkw.3	chr22	n/a	uc002zkw.3	DQ573684	
uc002zkk.2	chr22	n/a	uc002zkk.2	DQ595048	
uc002zky.2	chr22	n/a	uc002zky.2	DQ590589	
uc021wke.1	chr22	n/a	uc021wke.1	DQ573684	
uc002zla.2	chr22	n/a	uc002zla.2	DQ573684	
uc021wkf.1	chr22	n/a	uc021wkf.1	DQ587539	
uc002zlb.3	chr22	n/a	uc002zlb.3	DQ582484	
uc021wkg.1	chr22	n/a	uc021wkg.1	DQ599820	
uc021wkh.1	chr22	n/a	uc021wkh.1	DQ590589	
uc021wki.1	chr22	n/a	uc021wki.1	DQ573684	
uc002zlc.2	chr22	n/a	uc002zlc.2	DQ573684	
uc021wkj.1	chr22	n/a	uc021wkj.1	DQ587539	
uc002zld.3	chr22	n/a	uc002zld.3	DQ582484	
uc021wkk.1	chr22	n/a	uc021wkk.1	DQ599820	
uc021wk1.1	chr22	n/a	uc021wk1.1	DQ590589	
uc002zlf.2	chr22	n/a	uc002zlf.2	P704P	
uc002zlg.1	chr22	n/a	uc002zlg.1	POTEH	
uc002zlh.1	chr22	n/a	uc002zlh.1	POTEH	
uc010gqp.2	chr22	Q6S545	uc010gqp.2	POTEH	NM_001136213
uc002zlj.1	chr22	Q6S545	uc002zlj.1	POTEH	
uc002zlk.3	chr22	n/a	uc002zlk.3	P712P	
uc011agd.2	chr22	Q8NG94	GO:0001584,GO:0004871,GO:0004872,GO:0004930,GO:0004984,GO:0005886,GO:0007165,GO:0007186,GO:0007608,GO:001		
uc002zlo.1	chr22	n/a	uc002zlo.1	DQ571479	
uc002zlp.1	chr22	Q96SF2	GO:0000166,GO:0005515,GO:0005524,GO:0006457,GO:0044267,GO:0051082,	uc002zlp.1	CCTBL2 NM_014406
uc010gqq.3	chr22	n/a	uc010gqq.3	TPTEP1	
uc002z1a.4	chr22	n/a	uc002z1a.4	TPTEP1	

Looking at Biotypes and Patches

It's useful to be able to see the differences between the annotation in the genome browsers, and so here are some examples of how to find out biotypes and also how to view GRC patches in genomes.

Worked example 1:

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

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



STEP 3:
Search for gene symbol ABO

Search all categories ▾ ABO e.g. MRPS26 or AL035460.15

This Release

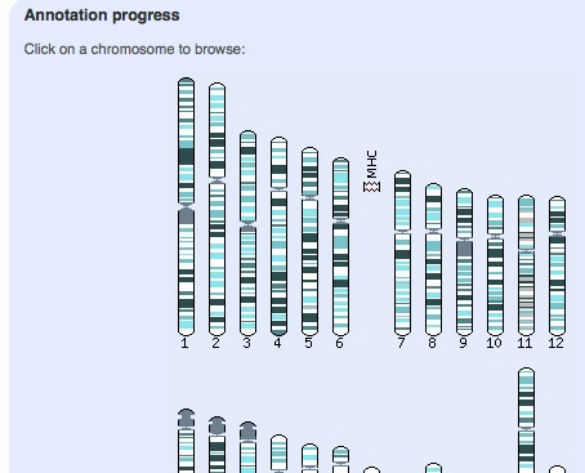
Release Date	22nd Sept 2014
Datafreeze Date	20th June 2014
Havana Update Date	25 November 2014
As used in	GENCODE 21; Ensembl 77 genebuild
Reference Assembly	GRCh38
Annotation Method	Complete first pass Manual Annotation

- Information and statistics
- Release 58 news
- Download our datasets (FTP)
- Go to Ensembl Human homepage

Example gene: Pax6, INS, FOXP2, BRCA2, DMD, ssh

Example transcript

Example region



Current selection:

< all Species

Only searching Human

Restrict category to:

Gene 3

Transcript 2

Per page:

10 25 50 100

Layout:

Standard Table

Tip:

If you have a search term with non alphanumeric characters you may get better results by enclosing the whole term in double quotes, for example "BRCA2-001".

Only searching Human ▾ ABO

5 results match ABO when restricted to species: Human

- ABO (Human Havana Gene)**
OTTHUMG00000020872 9:133250401-133275201:-1
ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) Havana annotation
Location • Sequence
- ABO-001 (Human Havana Transcript)**
OTTHUMT00000054907 9:133250401-133275201:-1
ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) Havana annotation
Location • cDNA seq. • Protein
- LOC401913-001 (Human Havana Transcript)**
OTTHUMT000000469665 19:34866352-34866919:-1
ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) pseudogene Havana annotation
Location • cDNA seq. • Protein
- LOC401913 (Human Havana Gene)**
OTTHUMG00000185147 19:34866352-34866919:1
ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) pseudogene Havana annotation
Location • Sequence
- HCG19P (Human Havana Alternate sequence Gene)**
OTTHUMG00000004819 6-QBL:1573903-1574536:-1
HLA complex group 19 pseudogene . Havana annotation Not a Primary Assembly Gene
Location • Sequence

STEP 4:
As this has search is by text there are 5 choices. The description explains how they differ. Click on the GeneID link of the top hit.

Human (VEGA57) Location: 9:133,250,401-133,275,201 Gene: ABO Transcript: ABO-001

Gene-based displays

- Summary
- Splice variants (1)
- Transcript comparison
- Supporting evidence
- Sequence
- External references
- Comparative Genomics
 - Genomic alignments
 - Orthologues
 - Alt. alleles
- External data
 - Personal annotation
- Other genome browsers
 - Ensembl

Gene: ABO OTTHUMG00000020872

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

Synonyms A3GALNT, A3GALT1

Location Chromosome 9: 133,250,401-133,275,201 reverse strand.

INSDC coordinates chromosome:VEGA57:CM000671.2:133250401:133275201:1

Transcripts This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
ABO-001	OTTHUMT00000054907	6341	No protein	Processed transcript	-	

Summary

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>

Version & date Version 5, last modified on 28/02/2014 (Created on 11/12/2003)

Annotation Attributes reference genome error [Definitions]

Other assemblies This gene maps to 133,250,401-133,275,201 in GRCh38 (Ensembl) coordinates. [Jump to this stable ID in Ensembl](#)

Curation Method Manual annotation from Havana

Alternative genes Ensembl gene: ENSG00000175164

STEP 5:
Notice that there are no protein products for this gene. The annotation attributes section shows that there is a genome error.

Summary

Reverse strand 24.80 kb

Statistics Exons: 7 Transcript length: 6,341 bps

Class processed transcript [Definition]

Author This transcript was annotated by Havana

Version & date Version 5, last modified on 28/02/2014 (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-*O01 allele
ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase), ABO-*O02 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 133,250,401-133,275,201 in GRCh38 (Ensembl) coordinates. [Jump to this stable ID in Ensembl](#)

Alternative transcripts Ensembl transcript having exact match with Havana: ENST00000453660

Curation Method Manual annotation from Havana

STEP 6:
Click on the transcript ID. The remarks field contains more detailed information.

STEP 7:
Click onto the location tab. You can then jump into Ensembl by clicking on the link at the side.

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 8:
The Ensembl tracks show 2 coding transcripts. Click on these to bring up the transcript view.

Transcript: ABO-202 ENST00000611156

Description ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) [Source:HGNC Symbol;Acc:HGNC:79]

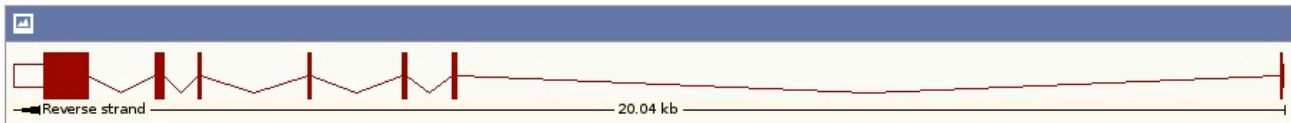
Synonyms A3GALNT, A3GALT1

Location [Chromosome 9: 133,255,176-133,275,214](#) reverse strand.

Gene This transcript is a product of gene [ENSG00000175164](#)
 This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
ABO-202	ENST00000611156	1577	353 aa	Protein coding	-	NM_020469 NP_065202	TSL:5 GENCODE basic APPRIS CI
ABO-201	ENST00000538324	1147	373 aa	Protein coding	-	-	TSL:5 GENCODE basic APPRIS CI1
ABO-001	ENST00000453660	6341	No protein	Processed transcript	-	-	TSL:1

Summary



Statistics Exons: 8 Coding exons: 8 Transcript length: 1,577 bps Translation length: 353 residues

Uniprot This transcript corresponds to the following Uniprot identifiers: [P16442](#)

Transcript Support Level 5

Ensembl version ENST00000611156.2

Type Known protein coding

Prediction Method Annotation produced by the Ensembl [genebuild](#).

Frameshift introns Frameshift introns occur at intron number(s) 6.

GENCODE basic gene This transcript is a member of the Gencode basic gene set.

STEP 9:
 There is a section called frameshift introns. Click on this to bring up the definition. Clicking on the number (6) will take you to the exons view. You can see the gt that was inserted to keep the coding frame.

Transcript Support Level 5

Ensembl version

Type

Prediction Method [genebuild](#).

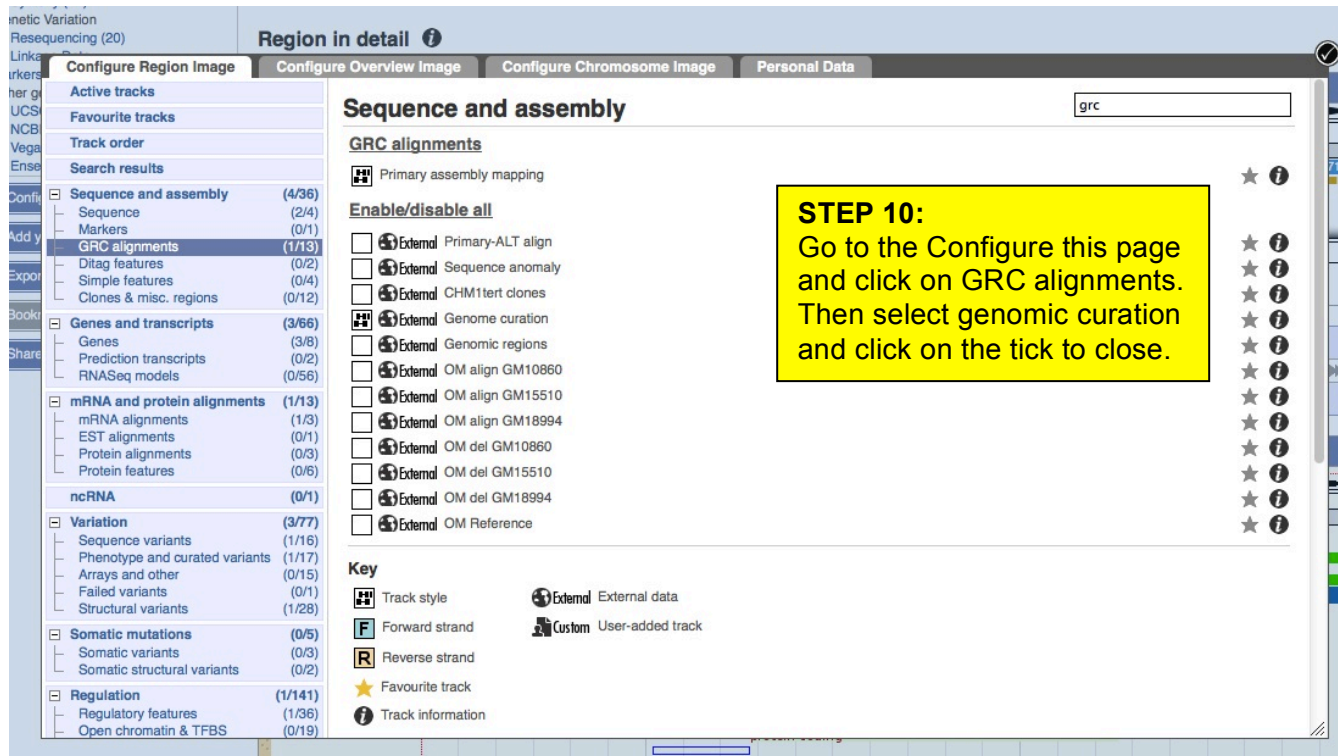
Frameshift introns Frameshift introns occur at intron number(s) 6.

GENCODE basic gene This transcript is a member of the Gencode basic

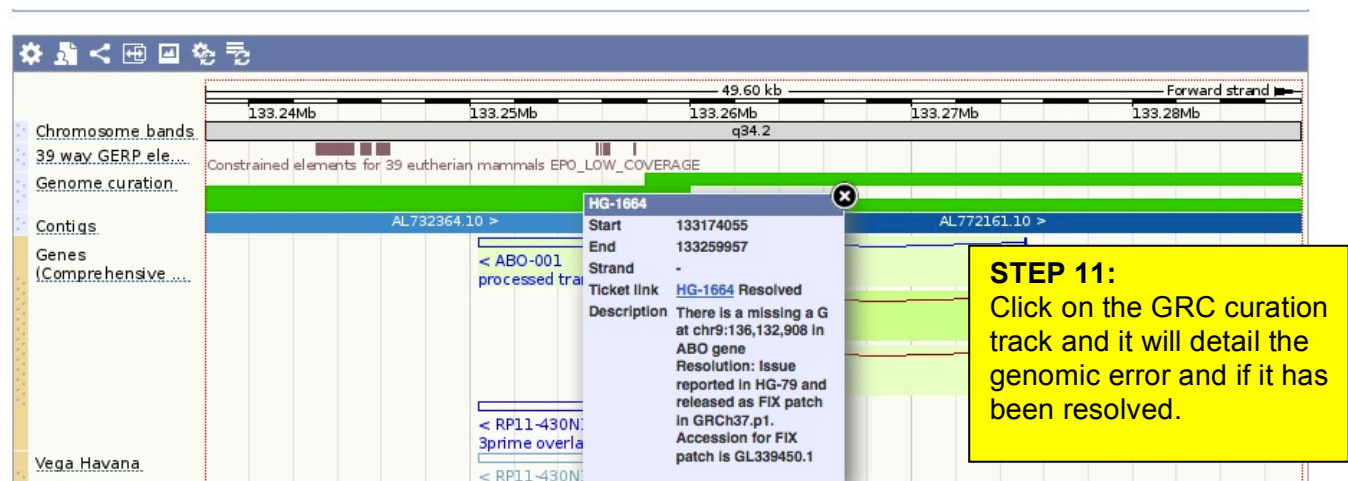
Frameshift introns are the length of 1, 2, 4, or 5 basepairs. They are introduced by the Ensembl genebuild in order to fit the cDNA sequence to the genome.

Intron 5-6	133,258,096	133,257,543		554	gtgagtaagttactgacactgaaa.....accgcacgcctctctccatgtgca
6 ENSE00003742406	133,257,542	133,257,524	2	0	19 TAGGAAGGATGCTCTCGTG
Intron 6-7	133,257,523	133,257,522		2	gt
7 ENSE00003725550	133,257,521	133,257,409	0	2	113 ACCCCTTGGCTGGCTCCCATTTGTCTGGGAGGGCACATTCAACATCGACATCCTCAACGACAGTTCAGGCTCCAGAACACCACCATTTGGGTTAAGTGTGTTTGCCATCAAGAA
Intron 7-8	133,257,408	133,256,357		1,052	gtaagtcagtgagtgccgagggg.....cagccccgtccgctgccttgca

You can view the GRC report in Ensembl to explain what the genomic problem is. In order to do this you need to configure the view.



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



Worked Example 2:

New and updated annotation is made available on a weekly basis by means of the Vega update. This means that annotation is available very quickly between Vega releases.

STEP 1:
Search for a gene ID:
OTTHUMG00000186331

STEP 2:
Click on the link to the updated annotation.

Gene: RP11-2E11.10 OTTHUMG00000186331

Updated annotation available
There is updated annotation for this gene available [here](#).

Description novel transcript antisense to MEST
Location Chromosome 7: 130,486,042-130,486,183 reverse strand.
INSDC coordinates chromosome:VEGA57:CM000669.2:130486042:130486183:1
Transcripts This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
RP11-2E11.10-001	OTTHUMT00000472944	142	No protein	Antisense	-	

Summary ⓘ

Curated Locus RP11-2E11.10 (Vega gene)
Gene type Novel antisense [Definition]
Author This gene was annotated by Havana <vega@sanger.ac.uk>
Version & date Version 1, last modified on 28/02/2014 (Created on 28/06/2013)
Other assemblies This gene maps to 130,486,042-130,486,183 in GRCh38 (Ensembl) coordinates. [Jump to this stable ID in Ensembl](#)

The gene is an antisense transcript with one splice variant.

STEP 3:
The updated annotation has an official gene symbol.

Gene: MESTIT1 OTTHUMG00000186331

Vega update gene

This is a Havana update gene with newer annotation than the core Vega gene.

Description	MEST intronic transcript 1, antisense RNA
Synonyms	MEST-AS1, NCRNA00040
Location	Chromosome 7: 130,486,042-130,491,033 reverse strand.
INSDC coordinates	chromosome:VEGA57:CM000669.2:130486042:130491033:1
Transcripts	

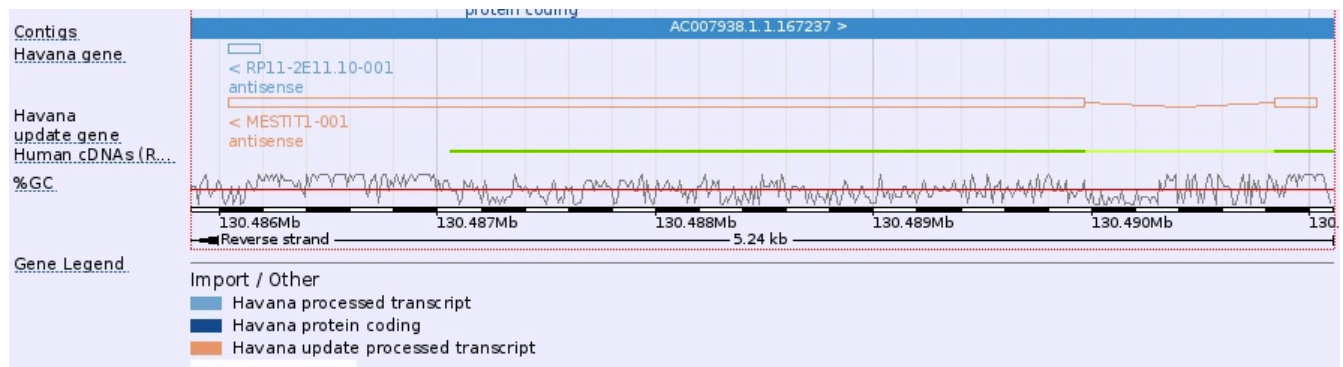
This gene has 1 transcript (splice variant) [Hide transcript table](#)

Show/hide columns		Filter				
Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
MESTIT1-001	OTTHUMT00000472944	4118	No protein	Antisense	-	

Summary

Curated Locus	MESTIT1 (HGNC Symbol)
Synonyms	MEST-AS1, NCRNA00040 [To view all genes linked to the name click here.]
Gene type	Novel Processed transcript [Definition]

In location view the update track is shown in brown:



Genes in Vega update may be new annotation or updated annotation.

Havana Update

The **Havana Update** gene set presents updates to annotation outside of the regular release schedule.

- [Further information.](#)
- [List of updated genes.](#)

Havana Human update genes (25 November 2014)

Gene Name	Biotype	Vega ID	Chromosome	Location	Modified date	New / Updated
RP11-278L15.7	processed_pseudogene	OTTHUMG00000190905	3	149463779-149464114	2014-11-24	new
RP11-1260E13.1	processed_transcript	OTTHUMG00000190904	HSCHR17_2_CTG1	80185-81246	2014-11-21	new
RP11-1260E13.4	processed_transcript	OTTHUMG00000190903	HSCHR17_2_CTG1	69150-71433	2014-11-21	new
RP11-1260E13.3	processed_transcript	OTTHUMG00000190902	HSCHR17_2_CTG1	45756-47901	2014-11-21	new
RP11-1260E13.2	processed_transcript	OTTHUMG00000190901	HSCHR17_2_CTG1	45057-50000	2014-11-21	new
RPH3AL	protein_coding	OTTHUMG00000190900	HSCHR17_2_CTG1	16384-122925	2014-11-21	new
KIAA0125	processed_transcript	OTTHUMG00000190899	HSCHR14_2_CTG1	295748-400411	2014-11-21	new

Worked Example 3:

Loss of function (LoF) transcripts are annotated for the predicted functional effects caused by single nucleotide variations. These originated from the pilot 1000 genomes project and are shown as a separate track. This work has been published by MacArthur et al (Science. 2012 Feb 17;335(6070):823 PMID: 22344438).

Search for the CNKSR1 gene in human Vega:



Only searching Human Gene ▾ **CNKSR1**

3 results match **CNKSR1** when restricted to species: Human ✕ category: Gene ✕

CNKSR1 (Human Havana Gene)**OTTHUMG0000007541** 1:26177403-26189886:1Connector enhancer of kinase suppressor of Ras 1 . *Havana annotation*[Location](#) • [Sequence](#)**LOF:CNKSR1 (Human Havana Gene)****OTTHUMG00000183607** 1:26177490-26189882:1LOF transcript from: connector enhancer of kinase suppressor of Ras 1 . *Havana annotation*[Location](#) • [Sequence](#)**RP11-173D3.3 (Human Havana Gene)****OTTHUMG00000171715** 15:19899334-19899559:1Connector enhancer of kinase suppressor of Ras 1 (**CNKSR1**) pseudogene *Havana annotation*[Location](#) • [Sequence](#)**STEP 1:**

Click on the gene link of the top hit.

Gene: CNKSR1 OTTHUMG00000007541

Description connector enhancer of kinase suppressor of Ras 1
Synonyms CNK, CNK1, KSR
Location Chromosome 1: 26,177,403-26,189,886 forward strand.
INSDC coordinates chromosome:VEGA57:CM000663.2:26177403:26189886:1
Transcripts

This gene has 15 transcripts (splice variants) [Hide transcript table](#)

Show All entries		Show/hide columns (1 hidden)		Filter		
Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
CNKSR1-001	OTTHUMT00000019856	2625	713 aa	Protein coding	CCDS276	
CNKSR1-007	OTTHUMT00000089943	2538	720 aa	Protein coding	CCDS72732	
CNKSR1-003	OTTHUMT00000019858	2673	455 aa	Protein coding	-	
CNKSR1-009	OTTHUMT00000383211	1065	No protein	Artifact	-	
CNKSR1-002	OTTHUMT00000019857	2507	229 aa	Nonsense mediated decay	-	
CNKSR1-011	OTTHUMT00000383212	1200	312 aa	Nonsense mediated decay	-	
CNKSR1-008	OTTHUMT00000089944	865	63 aa	Nonsense mediated decay	-	
CNKSR1-015	OTTHUMT00000383217	842	78 aa	Nonsense mediated decay	-	CDS 5' incomplete
CNKSR1-010	OTTHUMT00000383209	666	No protein	Processed transcript	-	
CNKSR1-014	OTTHUMT00000383216	553	No protein	Processed transcript	-	
CNKSR1-006	OTTHUMT00000019861	642	No protein	Retained intron	-	
CNKSR1-004	OTTHUMT00000019859	624	No protein	Retained intron	-	
CNKSR1-013	OTTHUMT00000383215	581	No protein	Retained intron	-	
CNKSR1-012	OTTHUMT00000383214	562	No protein	Retained intron	-	
CNKSR1-005	OTTHUMT00000019860	467	No protein	Retained intron	-	

The Vega annotation for this gene has 15 transcripts, of which 7 are protein coding.

Gene: LOF:CNKSR1 OTTHUMG00000183607

Loss of Function Variant

This gene is not standard Havana annotation, it is a Loss Of Function variant.

Description LOF transcript from: connector enhancer of kinase sup
Synonyms CNK, CNK1, KSR
Location Chromosome 1: 26,177,490-26,189,882 forward strand
INSDC coordinates chromosome:VEGA57:CM000663.2:26177490:26189882:1
Transcripts

STEP 2:
Click on the gene link of the LOF hit.

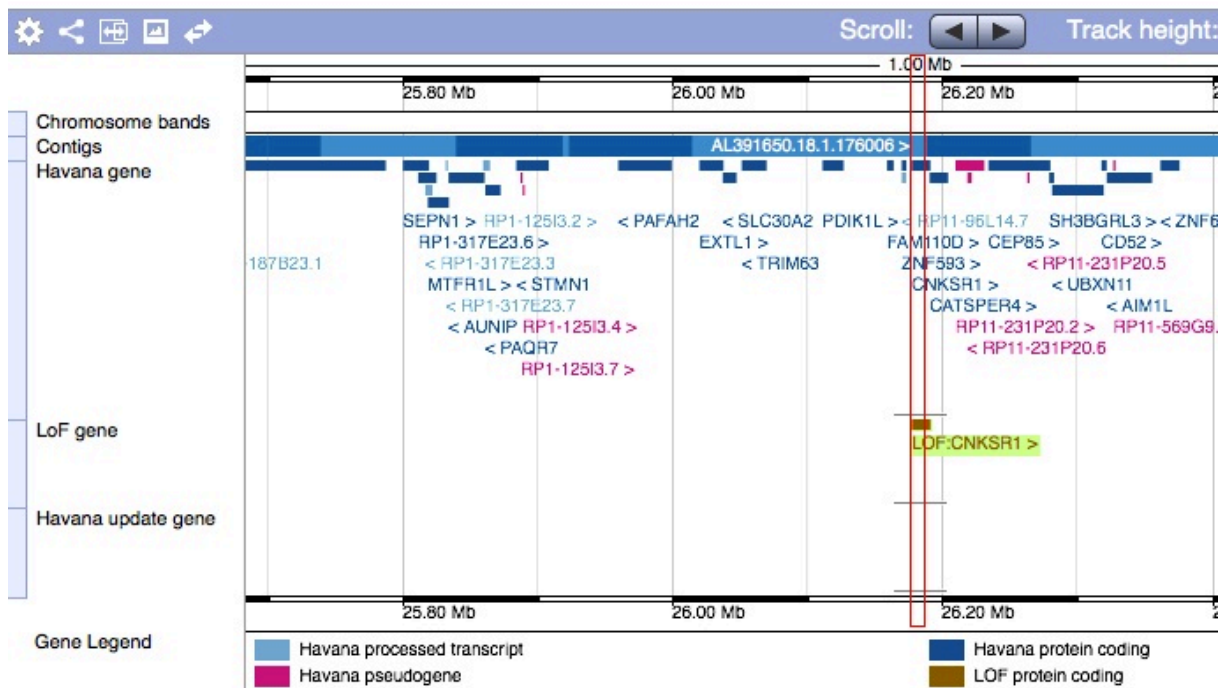
This gene has 2 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
LOF:CNKSR1-001	OTTHUMT00000383213	2507	180 aa	Nonsense mediated decay	-	
LOF:CNKSR1-002	OTTHUMT00000383210	666	169 aa	Nonsense mediated decay	-	

Summary

Curated Locus LOF:CNKSR1 (HGNC Symbol)
Gene type Putative protein coding [Definition]
Author This gene was annotated by Havana <vega@sanger.ac.uk>
Version & date Version 1, last modified on 27/02/2014 (Created on 16/08/2012)
Other assemblies This gene maps to 26,177,490-26,189,882 in GRCh38 (Ensembl) coordinates. Stable ID not present in Ensembl
Curation Method Transcripts annotated with the functional effects of Loss-of-function (LoF) variants identified in the 1000 Genomes Pilot Projects

There are two NMD transcripts which arise from SNVs from the 1000 genomes project, which could potentially code for 180 and 169 aa proteins. The LoF gene track can be viewed in Vega.



Tasks

1.
Search for the FCGR2C 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?

4.
Search for the gene ID OTTHUMG00000187111. Has the gene been updated?
What has changed?

5.
Search for the PLA2G2C gene. What extra information and gene tracks are available?

Answers:

1.

The FCGR2C gene has 10 variants in Vega. One of these is protein coding but there is a stop codon in the middle of the protein. This is because there is a SNP/DIP in this region of the reference genome that stops the gene from coding by introducing a stop codon. This polymorphism is known 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 57 brings up 19 entries. This is a simple text search that looks for the text string HERC2, so these are also brought up by the search.

There are 2 protein coding gene entries, one on the reference genome and one in an alternative assembly (HSCHR15_1_CTG8).

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 an alternative assembly (HSCHR15_4_CTG8).

4.

The gene has been updated with a new symbol LINC00540 and an extra splice variant.

5.

The PLA2G2C gene has a LoF gene that shows a truncated NMD protein of 32 aa.

