Module 2: The Vega and UCSC Genome Browsers

Aims

- Explain why it can be useful to look at the whole genome.
- Discuss how genes and other features can be predicted and displayed.
- Using Vega, and UCSC demonstrate some of the features and applications of genome browsers.
- Examples (include location and structure of a known gene and its products; information about a defined chromosomal region; convenient export of selected information).

Introduction

There has been an information explosion in molecular and genetic data for many organisms including full genome sequences, corresponding gene annotations, expanding transcription libraries and expression data, and better characterization of intra- and inter-species variation and conservation. 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 features of particular chromosomal regions
- Investigate specific genes as well as collections of genes
- Search for locations of sequences and markers
- Retrieve annotation information for specific regions or genome-wide
- View your own data in context of other annotations
- Compare one genome to genomes of other species

In addition to the genome sequence itself, browsers attempt to show the location and structure of numerous and diverse types of annotations such as genes. Importantly, contextual information such as GC content, locations of repetitive elements, evolutionary conservation, and regulatory information be viewed along with a gene of interest. Variation in the primary sequence of genes in terms of single nucleotide polymorphisms

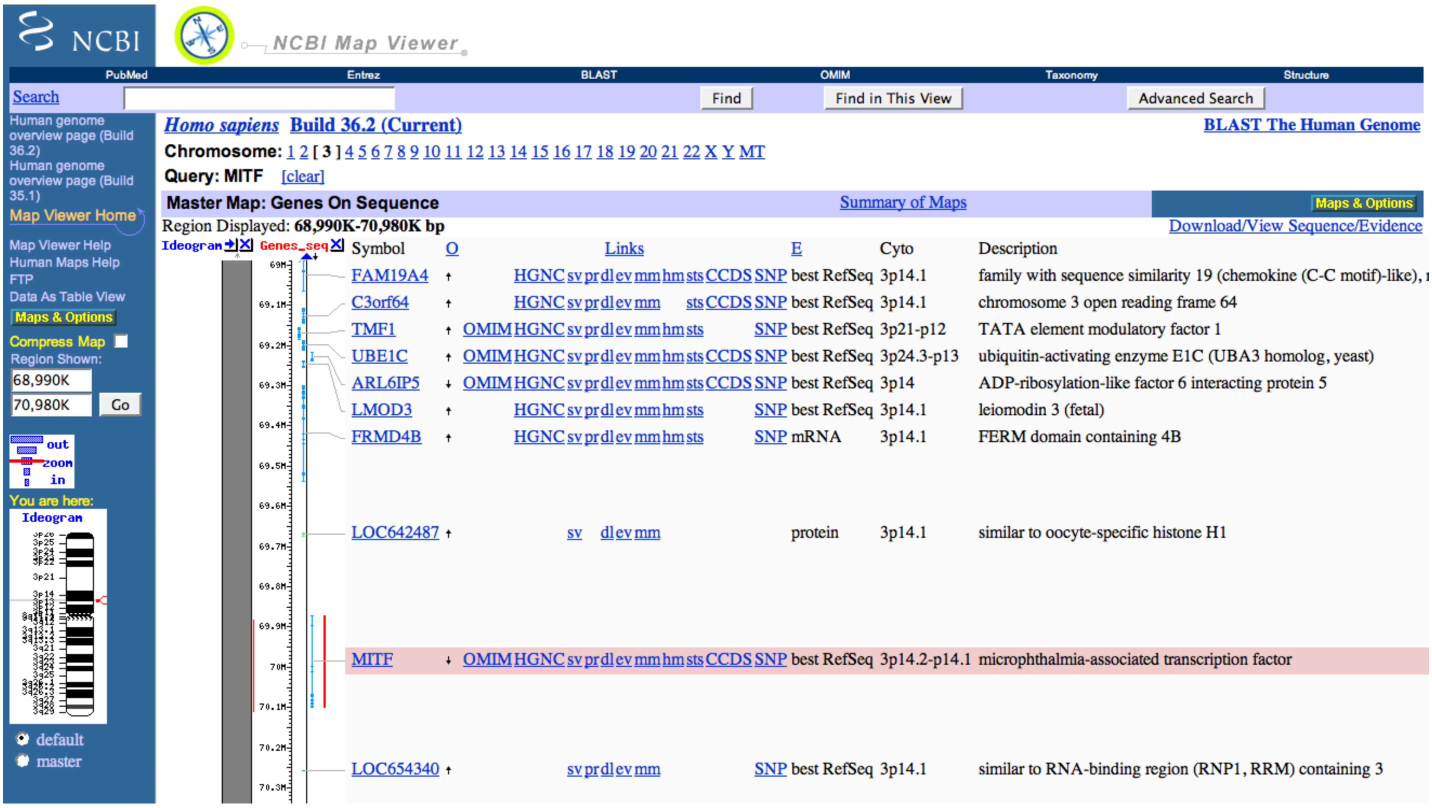
(SNPs) and larger structural variations can be assessed as well as transcriptional variation represented as alternatively spliced isoforms. Mapping information such as locations of sequence-tagged site (STS) markers, cytogenetic bands, and BAC and fosmid clones are available.

Human Genome Browsers

- **Map Viewer** – maintained by the NCBI
  http://www.ncbi.nih.gov/mapview/
- **Ensembl** – maintained by EBI / Sanger Institute http://ensembl.org/
- **Human Genome Browser** – maintained by UCSC http://genome.ucsc.edu

NCBI Map Viewer

http://www.ncbi.nlm.nih.gov/mapview/



Homo sapiens Build 36.2 (Current)

Chromosome: 1 2 [3] 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Query: MITF [clear]

Master Map: Genes On Sequence

Region Displayed: 68,990K-70,980K bp

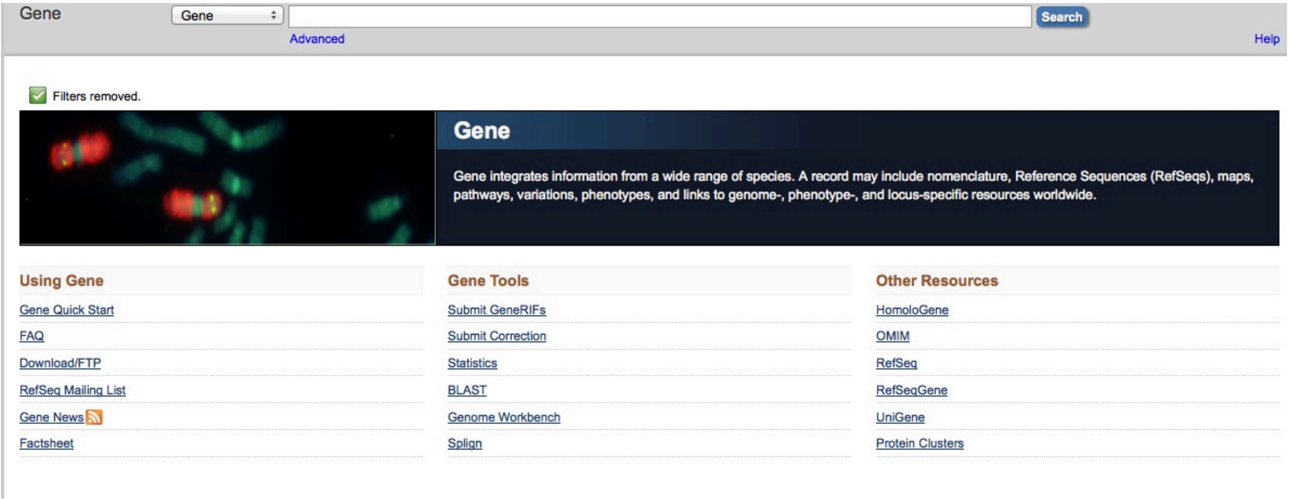
Symbol	O	Links	E	Cyto	Description
FAM19A4	↑	HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14.1	family with sequence similarity 19 (chemokine (C-C motif)-like),
C3orf64	↑	HGNC sv pr dl ev mm sts CCDS SNP	best RefSeq	3p14.1	chromosome 3 open reading frame 64
TMF1	↑	OMIM HGNC sv pr dl ev mm hm sts SNP	best RefSeq	3p21-p12	TATA element modulatory factor 1
UBE1C	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p24.3-p13	ubiquitin-activating enzyme E1C (UBA3 homolog, yeast)
ARL6IP5	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14	ADP-ribosylation-like factor 6 interacting protein 5
LMOD3	↑	HGNC sv pr dl ev mm hm sts SNP	best RefSeq	3p14.1	leiomodin 3 (fetal)
FRMD4B	↑	HGNC sv pr dl ev mm hm sts SNP	mRNA	3p14.1	FERM domain containing 4B
LOC642487	↑	sv dl ev mm	protein	3p14.1	similar to oocyte-specific histone H1
MITF	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14.2-p14.1	microphthalmia-associated transcription factor
LOC654340	↑	sv pr dl ev mm SNP	best RefSeq	3p14.1	similar to RNA-binding region (RNP1, RRM) containing 3

- Most unique design compared to UCSC and Ensembl
- Vertical view of chromosomes
- Excellent integration with other NCBI resources (Entrez Gene, HomoloGene, dbSNP, etc)
- Continually updated with new Genbank submissions

- Best "map" views of non-sequence maps
    - Genetic maps
    - RH maps
    - UniGene
    - OMIM Morbid
    - Mitelman breakpoints
    - Phenotype
- Includes all assemblies including Celera (Venter), Watson, alternate haplotypes
- Includes plant and fungal genome sequences
- BLAST used for sequencing searching, including in non-assembled sequences
- Maps displayed controlled through "Maps & Options" window
- Links to other databases provided on main display page

NCBI Gene

Not strictly a browser this is an excellent gene-centric resource from NCBI and is highly recommended. It links through to all available NCBI resources and the results can be customised.



☑ Filters removed.

Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene	Gene Tools	Other Resources
Gene Quick Start	Submit GeneRIFs	HomoloGene
FAQ	Submit Correction	OMIM
Download/FTP	Statistics	RefSeq
RefSeq Mailing List	BLAST	RefSeqGene
Gene News	Genome Workbench	UniGene
Factsheet	Splign	Protein Clusters

Other Genome Browsers

These are not the only genome browsers available, but are simply the primary ones for the human genome. Many other organism specific browsers exist including the Sacchromyces Genome Database (SGD), WormBase, FlyBase, the Rat Genome Database (RGD), and the Mouse Genome Informatics (MGI) database, to name a few.

Data retrieval and data mining

Genomic annotation data, due to its complexity and volume, does not lend itself to easy access. Presenting it on a web site is important, but so is providing simple but flexible ways to select and retrieve all or specific sets of data. NCBI uses the Entrez query system, UCSC provides this through full genome database downloads and their Table Browser, and Ensembl has developed a tool called BioMart.

The annotation data that is provided through these websites are stored in relational databases designed independently by each resource. The underlying sequence on which annotations are made is exactly the same at all sites and is distributed by NCBI. Specific annotations can and do vary on each site due to variations in methods used by each to create the annotations. For example, the alignment of a specific mRNA sequence may not be exactly the same due to the use of different alignment programs or parameter settings within the program. Some annotations are shared by all browsers such as the locations of cytogenetic bands (for human).

Displaying your own data

The UCSC Genome Browser and the Ensembl browser provide the ability to view genome annotations created by you within their browsers. Simply organizing your data into one of several types of file formats and uploading it into the browsers allow you to privately view it as if it were part one of the provided annotations. You can also use your data within the UCSC Table Browser to filter and download information from other annotations.

ENCODE (ENCycolopedia of DNA Elements) Project

One of the successor projects to the Human Genome Project is the ENCODE Project, an effort to define all of the functional DNA elements in the genome. This includes locations of all genes including protein coding, non-coding, RNA, and pseudogenes with multiple splice forms, transcription factor binding sites, histone modifications, and chromatin structure. The pilot phase of the project was completed in Spring of 2007 that focused on 1% (30Mb) of the genome, and the scale-up phase evaluating the whole genome is underway and near to completion.

In the scale-up phase, most of the groups are utilizing second-generation sequencing technologies, especially the Illumina sequencer. The short sequence tags produced by these machines in large quantities allow for the identification of regions of interest (i.e.

transcription factor binding site) and to better characterize transcripts. Other sequencers such as from 454, Applied Biolosystems (SOLiD), and Helicos are available with varying strengths and weaknesses.

The genome-wide data from the ENCODE project is now available in Ensembl and UCSC.

CCDS

Vega 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), the Hugo Genome Nomenclature Committee (HGNC) and Mouse Genome Informatics (MGI).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 HGNC (for human) and MGI (for mouse). 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.

The Genome Reference Consortium

This is a joint initiative between the Wellcome Trust Sanger Institute, the European Bioinformatics Institute, The Genome Centre at Washington University and the NCBI. The goal is to correct regions in the genome that are currently misrepresented, to close as many gaps as possible, to produce alternative assemblies of structurally variant loci where necessary and a means to report loci in need of review. The GRC is in place for the human, mouse and zebrafish genomes.

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml

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. Vega displays complete chromosomes and specific regions of interest. Grey shading indicates annotation status, with light grey showing partially annotated regions and dark grey showing regions with no annotation. Currently the available species are human, mouse, rat, zebrafish, pig, tammar wallaby, gorilla, tasmanian devil, chimpanzee, and dog.

Vega genes are displayed in the UCSC Genome Browser as part of the GENCODE geneset and in Ensembl they are part of the merged gene-set. They may also be viewed as a separate track in both browsers.

Human

The following groups have contributed to the annotation of whole human chromosomes:

* The Havana group (chr. 1, 2, 6, 7, 9, 10, 13, 20, 21, 22, X, Y) and Collins *et al.* (chr. 22) at the Wellcome Trust Sanger Insitute.

* Hillier et al. (chr. 7) at the Washington University Genome Center

* Genoscope (chr. 14) at CNRS

* The DOE Joint Genome Institute (chr. 16, 19)

* The Broad Institute (chr. 8, 15, 17, 18)

* Baylor College of Medicine (chr. 3, 12)

* Genome Analysis Group (chr. X) at the Institute of Mol. Biotechnology

First pass manual annotation has now been completed for the whole human genome.

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 chromosomes and are also included in the Vega comparative analysis.

The human Leucocyte Receptor Complex (LRC) is located on chromosome 19q13.4 in a region spanning approximately 1.0 Mb. It comprises sets of genes encoding natural killer receptors, the Killer Immunoglobulin-like Receptors (KIR) as well as related immunoglobulin superfamily genes, with the latter including the Leucocyte Immunoglobulin-Like Receptors (LILR) and the Leucocyte-Associated Immunoglobulin-like Receptors (LAIR). The LRC region on chromosome 19 has been annotated in two haplotypes: COX and PGF_1. As for the MHC regions, these are shown as distinct chromosomes, for example 19-COX. They are included in the Vega comparative analysis.

Gorilla has a series of BACs that comprise the complete gorilla classical MHC region on chromosome 6, including part of the extended MHC. Separately, a small contig of chromosome 19 BACs from the same library were sequenced for the annotation of KIR (Killer Immunoglobulin-like Receptor) genes on chromosome 19. This contig contains the KIR genes, the KIR3DX1 ancestral gene and a large part of the rest of the LRC (Leukocyte Receptor Complex) cluster.

Pig has the full genome (Build 10.2), with over 1500 genes manually annotated as part of the Immune Response Annotation Group (IRAG). This community annotation effort involved over 30 annotators worldwide targeting genes involved in immune response and was QC'd by the Havana team to ensure high quality annotation. This annotation is marked as IRAG annotation in Vega. The MHC (swine leukocyte antigen complex, SLA) region on chromosome 7, from Large White Boar. SLA molecules are of interest for their potential role in xenotransplantation reactions. An 8Mb region of chromosome 17 as is syntenic to human chromosome 20q13.13-q13.33 and mouse chromosome 2 (167.44Mb-178.12Mb). Amongst other genes of interest this region contains the GNAS complex locus, which exhibits a highly complex imprinted expression pattern. Also a region of ~300kb of chromosome 6 (two BAC clones) has been annotated due to its orthology with part of the leukocyte receptor complex (LRC) on human 19q13.4.

WTSI has the sequenced the X and Y from pig and has manually annotated them. The pig genome build 10.2 is represented in Vega, plus the WTSI X and Y chromosomes which contain new clones and extra annotation compared to the build 10.2 chromosomes.

Wallaby has contigs and isolated BACs containing MHC related genes. The sequence contains Extended Class I and Class I and II regions, antigen processing genes, Class II DAA, DAB, DBA and DBB genes and Class II pseudogenes and olfactory receptor genes. The MHC of the wallaby is fragmented, with the primary gene cluster located on chromosome 2q and ten BACs containing class I genes located at six different chromosomal locations.

Dog shows the MHC (DLA) class II region on chromosome 12 from Doberman from five BAC clones.

Mouse

Mouse currently has full manual annotation of chromosomes 1, 2, 3, 4, 11 and X. Annotation of clones and loci of specific interest, spread throughout the genome, are also presented.
The DeL36H regions of chromosome 13 are shown, which exhibit synteny with part of human chromosome 6, particularly regions deleted in human syndromes and various disease loci. This means that the Del36H mice may be used to investigate certain human conditions. Candidate Insulin dependent diabetes (IDD) regions on chromosomes 1, 3, 4, 6, 11 and 17 have been annotated in both the CL57/BL6 reference strain and one or more of DIL NOD, CORI-29 NOD and 129 strains. Vega contains a comparative analysis of these regions in the different strains.
Havana is annotating several thousand genes as part of the European Conditional Mouse Mutagenesis Program (EUCOMM) and NIH Knockout Mouse Program (KOMP). These projects are designed to produce comprehensive libraries of mouse embryonic stem (ES) cells containing null mutations in every gene in the mouse genome. Manual annotation is needed to identify the exons required for correct translation of the longest transcript of the target loci. With this knowledge constructs can be designed for gene targeting experiments, one of the approaches used for knocking out target genes. Vega shows these exons as the 'knockout deletions' and it also contains 'Knockout genes' which are the artificial genes arising from the removal of these exons.

Rat has the full genome (currently build 6.0) with manual annotation of targeted regions and gene families that have been requested by the Rat Genome Database (RGD) and the rat community. There are currently 1650 genes in Vega together with a further 100 in the rat update track.

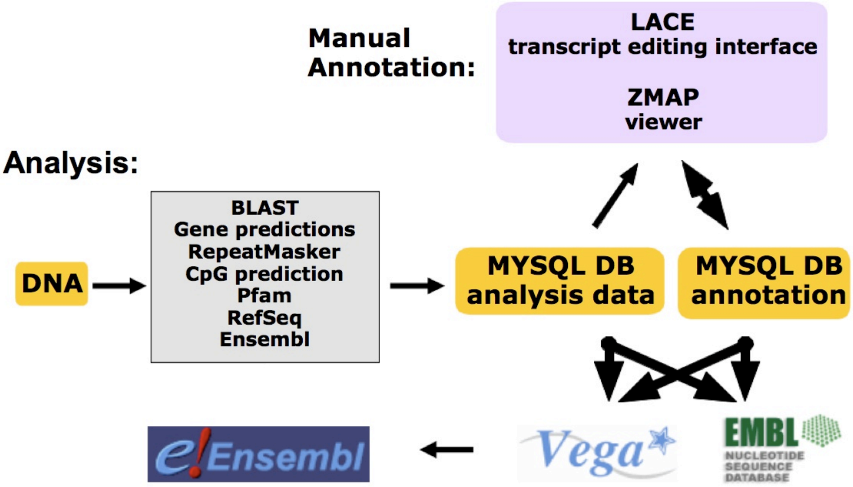
Zebrafish

Zebrafish has annotation for chromosomes 1, 2, 4, 5, 8, 9, 10, 13, 16, 18, 19, 20, 21, 22, 23 and 24 currently displayed. The genome is being sequenced and assembled in its entirety at WTSI.

CCDS

Vega 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), the HUGO Genome Nomenclature Committee (HGNC) and Mouse Genome Informatics (MGI). The aim of the project is to identify a core set of human and mouse 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.

Manual Genome Annotation



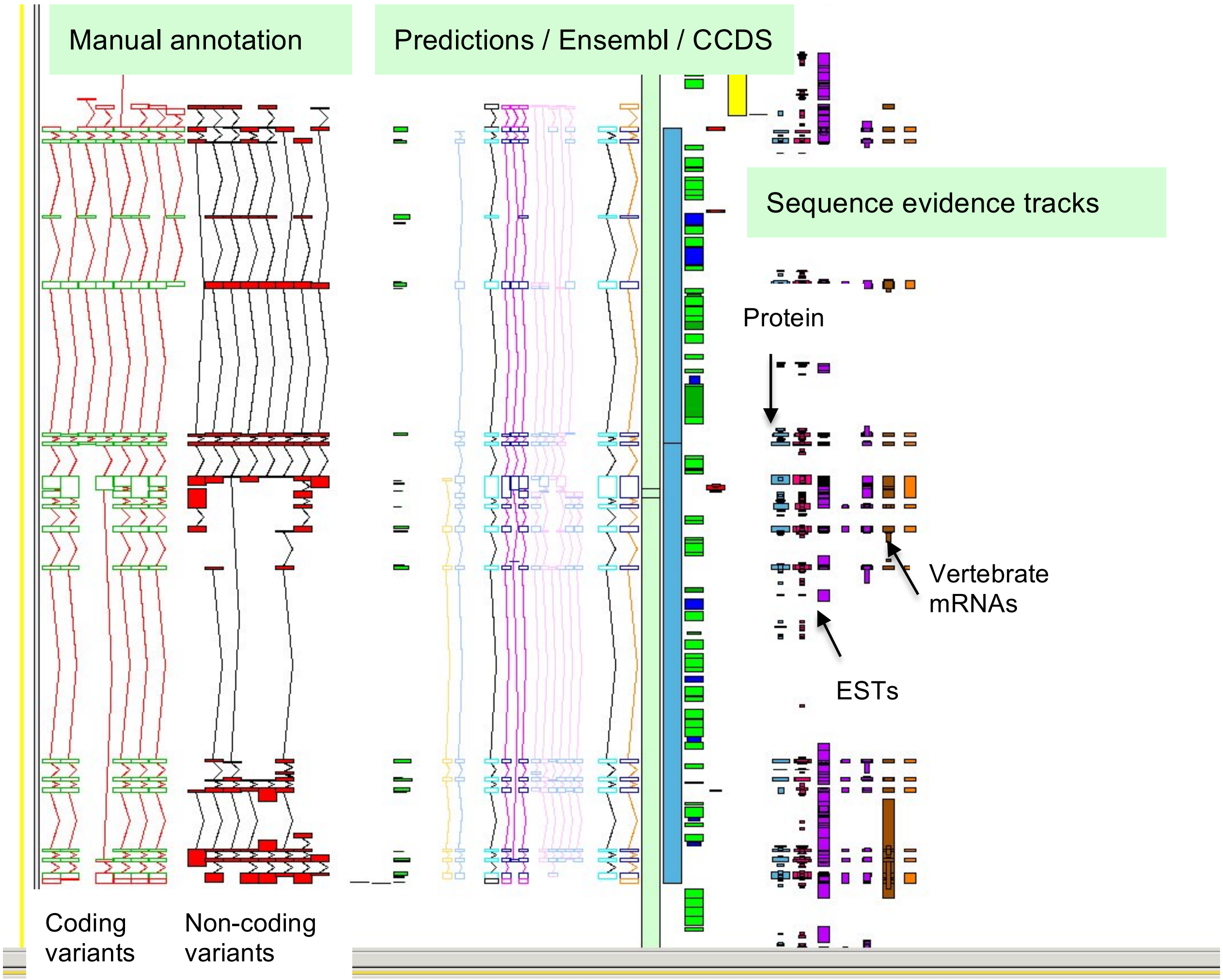
HAVANA (Human and Vertebrate Analysis and Annotation) group at the WTSI perform manual genome annotation. Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases (including cross-species) and a series of *ab initio* gene predictions (the analysis pipeline). Annotation is based on supporting evidence, which is external sequence such as ESTs, cDNAs and protein. There are multiple biotypes that reflect confidence levels and there are additional data sources included as DAS tracks (e.g. CAGE tags, RNAseq). The annotators then view this data through the Zmap viewer and perform manual annotation in the Otterlace transcript editing interface. The annotation is then saved 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.

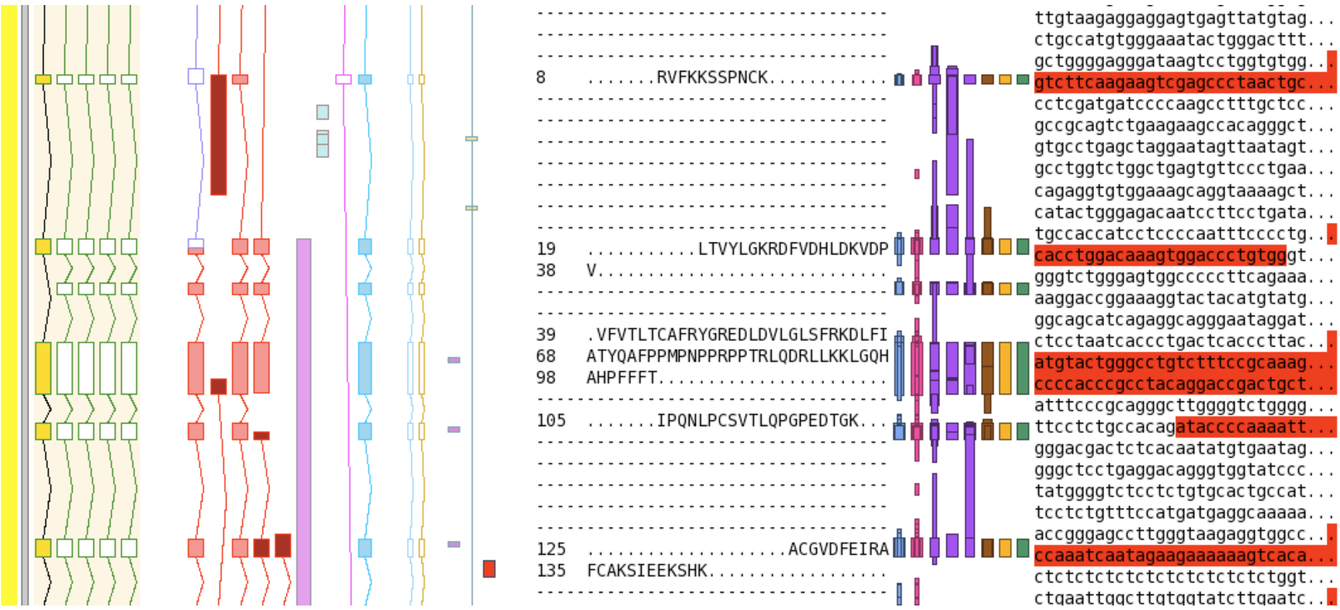
Vega may be browsed and searched in a similar way to Ensembl.

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).

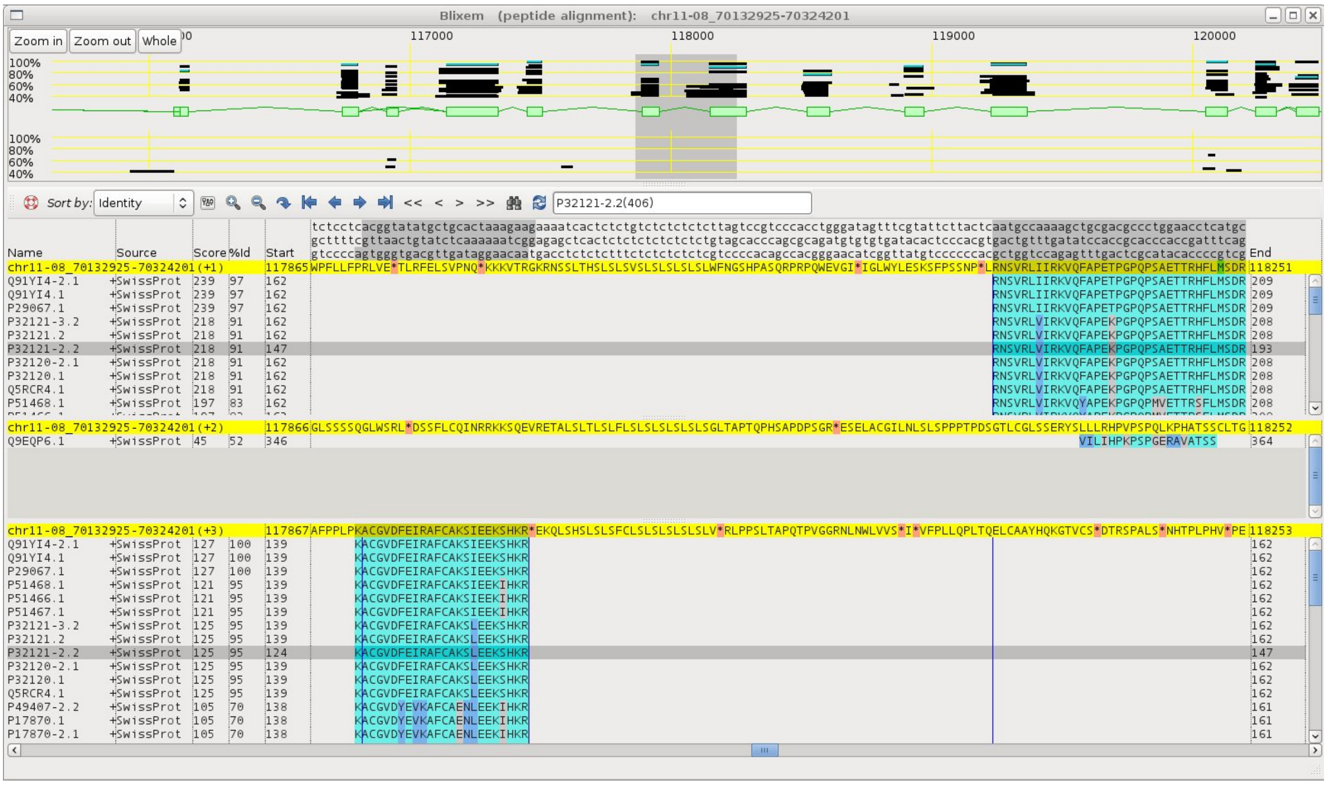


Otterlace manual annotation tool overview:



```
-----------------------------------           ttgtaagaggaggagtgagttatgtag...
-----------------------------------           ctgccatgtgggaaatactgggacttt...
-----------------------------------           gctggggagggataagtcctggtgtgg...
8     .......RVFKKSSPNCK...........           gtcttcaagaagtcgagccctaactgc...
-----------------------------------           cctcgatgatccccaagcctttgctcc...
-----------------------------------           gccgcagtctgaagaagccacagggct...
-----------------------------------           gtgcctgagctaggaatagttaatagt...
-----------------------------------           gcctggtctggctgagtgttccctgaa...
-----------------------------------           cagaggtgtggaaagcaggtaaaagct...
-----------------------------------           catactgggagacaatccttcctgata...
-----------------------------------           tgccaccatcctccccaatttcccctg...
19    ...........LTVYLGKRDFVDHLDKVDP           cacctggacaaagtggaccctgtgggt...
38    V.............................           gggtctgggagtggccccttcagaaa...
-----------------------------------           aaggaccggaaaggtactacatgtatg...
-----------------------------------           ggcagcatcagaggcagggaataggat...
39    .VFVTLTCAFRYGREDLDVLGLSFRKDLFI           ctcctaatcaccctgactcacccttac...
68    ATYQAFPPMPNPPRPPTRLQDRLLKKLGQH           atgtactgggcctgtctttccgcaaag...
98    AHPFFFT.......................           ccccacccgcctacaggaccgactgct...
-----------------------------------           atttcccgcagggcttggggtctgggg...
105   .......IPQNLPCSVTLQPGPEDTGK...           ttcctctgccacagataccccaaaatt...
-----------------------------------           gggacgactctcacaatatgtgaatag...
-----------------------------------           gggctcctgaggacagggtggtatccc...
-----------------------------------           tatggggtctcctctgtgcactgccat...
-----------------------------------           tcctctgtttccatgatgaggcaaaaa...
-----------------------------------           accgggagccttgggtaagaggtggcc...
125   ....................ACGVDFEIRA           ccaaatcaatagaagaaaaaagtcaca...
135   FCAKSIEEKSHK..................           ctctctctctctctctctctctctggt...
-----------------------------------           ctgaattggcttgtggtatcttgaatc...
```

The graphical interface can display numerous tracks (including DAS tracks), such as protein translations, genomic DNA, cDNA and ESTs, transcript model (see above). Multiple sequence alignments of protein and nucleotide sequences can be viewed in detail (see below).



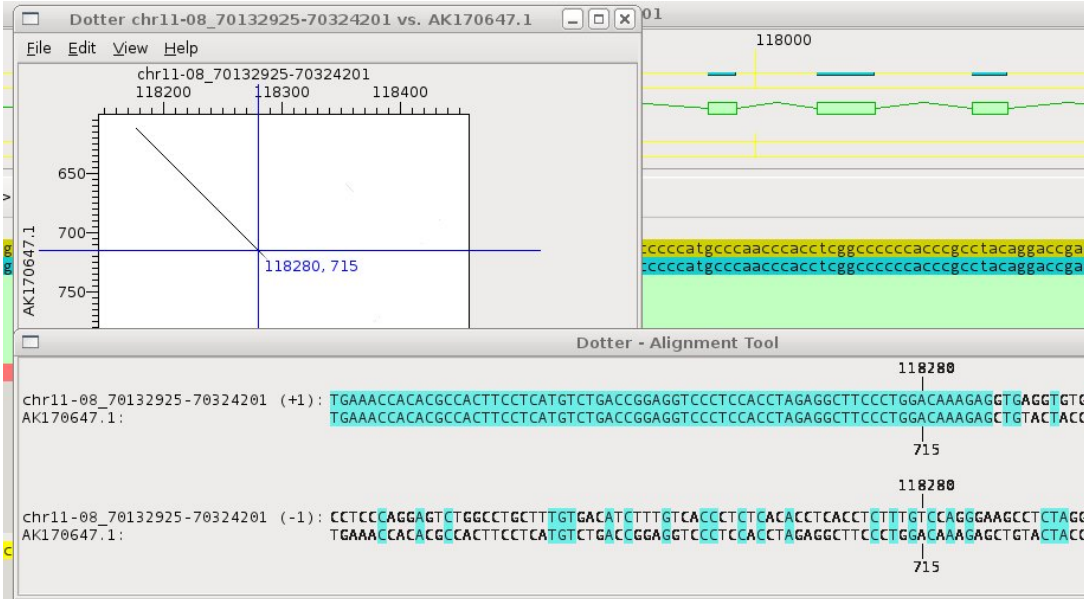
Blixem (peptide alignment): chr11-08_70132925-70324201

Sort by: Identity

P32121-2.2(406)

```
                                                          tctcctcacggtatatgctgcactaaagaagaaaatcactctctgtctctctctcttagtccgtcccacctgggatagtttcgtattcttactcaatgccaaaagctgcgacgccctggaacctcatgc
                                                          gcttttcgttaactgtatctcaaaaaatcggagagctcactctctctctctctctgtagcacccagcgcagatgtgtgtgatacactcccacgtgactgtttgatatccaccgcacccaccgatttcag
Name                       Source       Score %Id  Start  gtccccagtgggtgacgttgataggaacaatgacctctctctttctctctctctcgtccccacagccacgggaacatcggttatgtcccccacgctggtccagagtttgactcgcatacaccccgtcg End
chr11-08_70132925-70324201(+1)                     117865 WPFLLFPRLVE*TLRFELSVPNQ*KKKVTRGKRNSSLTHSLSLSVSLSLSLSLSLWFNGSHPASQRPRPQWEVGI*IGLWYLESKSFPSSNP*LRNSVRLIIRKVQFAPETPGPQPSAETTRHFLMSDR 118251
Q91YI4-2.1                 +SwissProt   239   97   162                                                                                           RNSVRLIIRKVQFAPETPGPQPSAETTRHFLMSDR 209
Q91YI4.1                   +SwissProt   239   97   162                                                                                           RNSVRLIIRKVQFAPETPGPQPSAETTRHFLMSDR 209
P29067.1                   +SwissProt   239   97   162                                                                                           RNSVRLIIRKVQFAPETPGPQPSAETTRHFLMSDR 209
P32121-3.2                 +SwissProt   218   91   162                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 208
P32121.2                   +SwissProt   218   91   162                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 208
P32121-2.2                 +SwissProt   218   91   147                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 193
P32120-2.1                 +SwissProt   218   91   162                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 208
P32120.1                   +SwissProt   218   91   162                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 208
Q5RCR4.1                   +SwissProt   218   91   162                                                                                           RNSVRLVIRKVQFAPEKPGPQPSAETTRHFLMSDR 208
P51468.1                   +SwissProt   197   83   162                                                                                           RNSVRLVIRKVQYAPEKPGPQPMVETTRSFLMSDR 208
chr11-08_70132925-70324201(+2)                     117866 GLSSSSQGLWSRL*DSSFLCQINRRKKSQEVRETALSLTLSLFLSLSLSLSLSLSGLTAPTQPHSAPDPSGR*ESELACGILNLSLSPPPTPDSGTLCGLSSERYSLLLRHPVPSPQLKPHATSSCLTG 118252
Q9EQP6.1                   +SwissProt   45    52   346                                                                                                       VILIHPKPSPGERAVATSS     364
chr11-08_70132925-70324201(+3)                     117867 AFPPLPKACGVDFEIRAFCAKSIEEKSHKR*EKQLSHSLSLSFCLSLSLSLSLSLV*RLPPSLTAPQTPVGGRNLNWLVVS*I*VFPLLQPLTQELCAAYHQKGTVCS*DTRSPALS*NHTPLPHV*PE 118253
Q91YI4-2.1                 +SwissProt   127   100  139          KACGVDFEIRAFCAKSIEEKSHKR                                                                                                    162
Q91YI4.1                   +SwissProt   127   100  139          KACGVDFEIRAFCAKSIEEKSHKR                                                                                                    162
P29067.1                   +SwissProt   127   100  139          KACGVDFEIRAFCAKSIEEKSHKR                                                                                                    162
P51468.1                   +SwissProt   121   95   139          KACGVDFEIRAFCAKSIEEKIHKR                                                                                                    162
P51466.1                   +SwissProt   121   95   139          KACGVDFEIRAFCAKSIEEKIHKR                                                                                                    162
P51467.1                   +SwissProt   121   95   139          KACGVDFEIRAFCAKSIEEKIHKR                                                                                                    162
P32121-3.2                 +SwissProt   125   95   139          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    162
P32121.2                   +SwissProt   125   95   139          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    162
P32121-2.2                 +SwissProt   125   95   124          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    147
P32120-2.1                 +SwissProt   125   95   139          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    162
P32120.1                   +SwissProt   125   95   139          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    162
Q5RCR4.1                   +SwissProt   125   95   139          KACGVDFEIRAFCAKSLEEKSHKR                                                                                                    162
P49407-2.2                 +SwissProt   105   70   138          KACGVDYEVKAFCAENLEEKIHKR                                                                                                    161
P17870.1                   +SwissProt   105   70   138          KACGVDYEVKAFCAENLEEKIHKR                                                                                                    161
P17870-2.1                 +SwissProt   105   70   138          KACGVDYEVKAFCAENLEEKIHKR                                                                                                    161
```

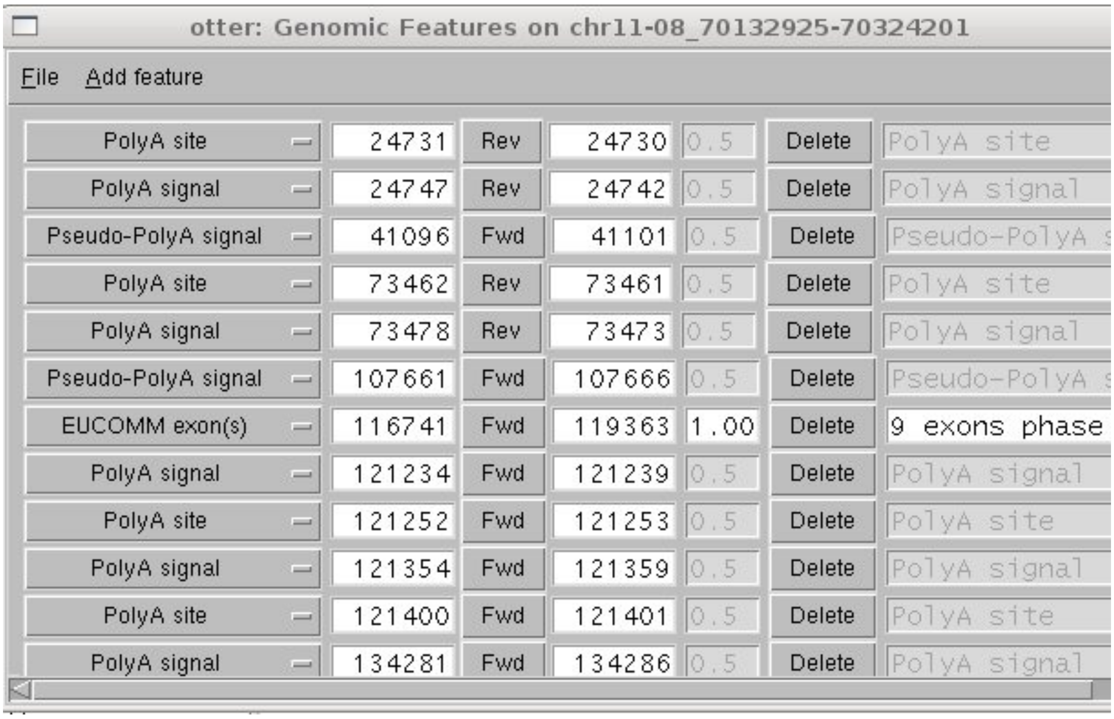
Pairwise sequence alignments of proteins and nucleotide sequences with transcript models or genomic sequence can be shown in detail (see below)



```
ccccatgcccaacccacctcggcccccacccgcctacaggaccga
ccccatgcccaacccacctcggcccccacccgcctacaggaccga
```

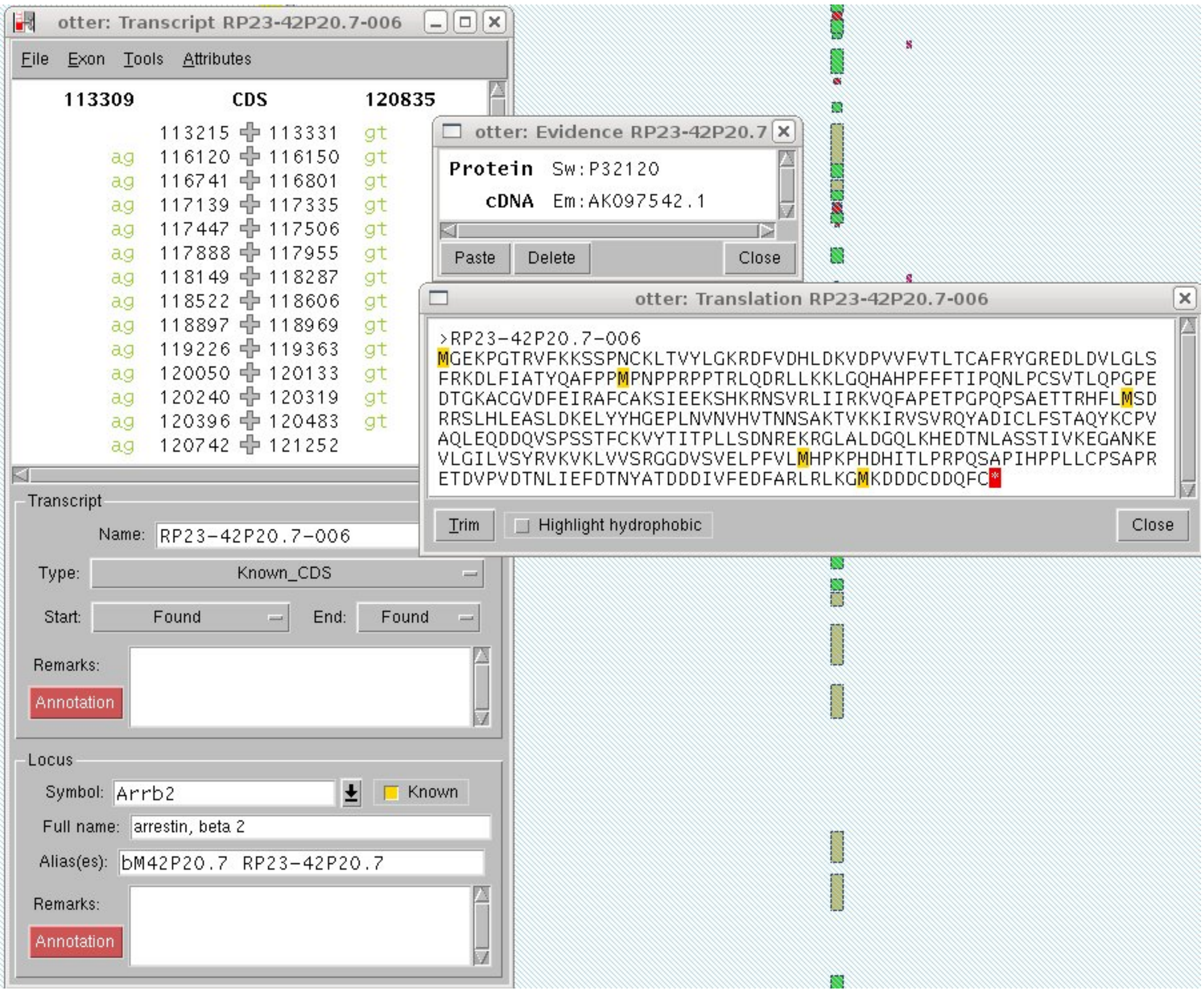
```
                                                                                              118280
                                                                                                 |
chr11-08_70132925-70324201 (+1): TGAAACCACACGCCACTTCCTCATGTCTGACCGGAGGTCCCTCCACCTAGAGGCTTCCCTGGACAAAGAGGTGAGGTGT
AK170647.1:                      TGAAACCACACGCCACTTCCTCATGTCTGACCGGAGGTCCCTCCACCTAGAGGCTTCCCTGGACAAAGAGCTGTACTAC
                                                                                                 |
                                                                                                715

                                                                                              118280
                                                                                                 |
chr11-08_70132925-70324201 (-1): CCTCCCAGGAGTCTGGCCTGCTTTGTGACATCTTTGTCACCCTCTCACACCTCACCTCTTTGTCCAGGGAAGCCTCTAG
AK170647.1:                      TGAAACCACACGCCACTTCCTCATGTCTGACCGGAGGTCCCTCCACCTAGAGGCTTCCCTGGACAAAGAGCTGTACTAC
                                                                                                 |
                                                                                                715
```

Genomic features such as polyA sites and signals and knock-out target exons for the mouse EUCOMM project are added (see below).



PolyA site	24731	Rev	24730	0.5	Delete	PolyA site
PolyA signal	24747	Rev	24742	0.5	Delete	PolyA signal
Pseudo-PolyA signal	41096	Fwd	41101	0.5	Delete	Pseudo-PolyA s
PolyA site	73462	Rev	73461	0.5	Delete	PolyA site
PolyA signal	73478	Rev	73473	0.5	Delete	PolyA signal
Pseudo-PolyA signal	107661	Fwd	107666	0.5	Delete	Pseudo-PolyA s
EUCOMM exon(s)	116741	Fwd	119363	1.00	Delete	9 exons phase
PolyA signal	121234	Fwd	121239	0.5	Delete	PolyA signal
PolyA site	121252	Fwd	121253	0.5	Delete	PolyA site
PolyA signal	121354	Fwd	121359	0.5	Delete	PolyA signal
PolyA site	121400	Fwd	121401	0.5	Delete	PolyA site
PolyA signal	134281	Fwd	134286	0.5	Delete	PolyA signal

The interface for building transcript models includes coordinates of exons, splice site sequence, transcript and locus names. Supporting evidence from the databases is added here and the translation of the CDS can be viewed with the methionine residues highlighted.


otter: Transcript RP23-42P20.7-006
File Exon Tools Attributes
otter: Evidence RP23-42P20.7
Paste
Delete
Close
otter: Translation RP23-42P20.7-006
Trim
□ Highlight hydrophobic
Close
Annotation
Annotation

```
     113309         CDS        120835
          113215 ✚ 113331   gt
     ag   116120 ✚ 116150   gt
     ag   116741 ✚ 116801   gt
     ag   117139 ✚ 117335   gt
     ag   117447 ✚ 117506   gt
     ag   117888 ✚ 117955   gt
     ag   118149 ✚ 118287   gt
     ag   118522 ✚ 118606   gt
     ag   118897 ✚ 118969   gt
     ag   119226 ✚ 119363   gt
     ag   120050 ✚ 120133   gt
     ag   120240 ✚ 120319   gt
     ag   120396 ✚ 120483   gt
     ag   120742 ✚ 121252
```

Transcript

Name: RP23-42P20.7-006

Type: Known_CDS

Start: Found End: Found

Remarks:

Locus

Symbol: Arrb2 Known

Full name: arrestin, beta 2

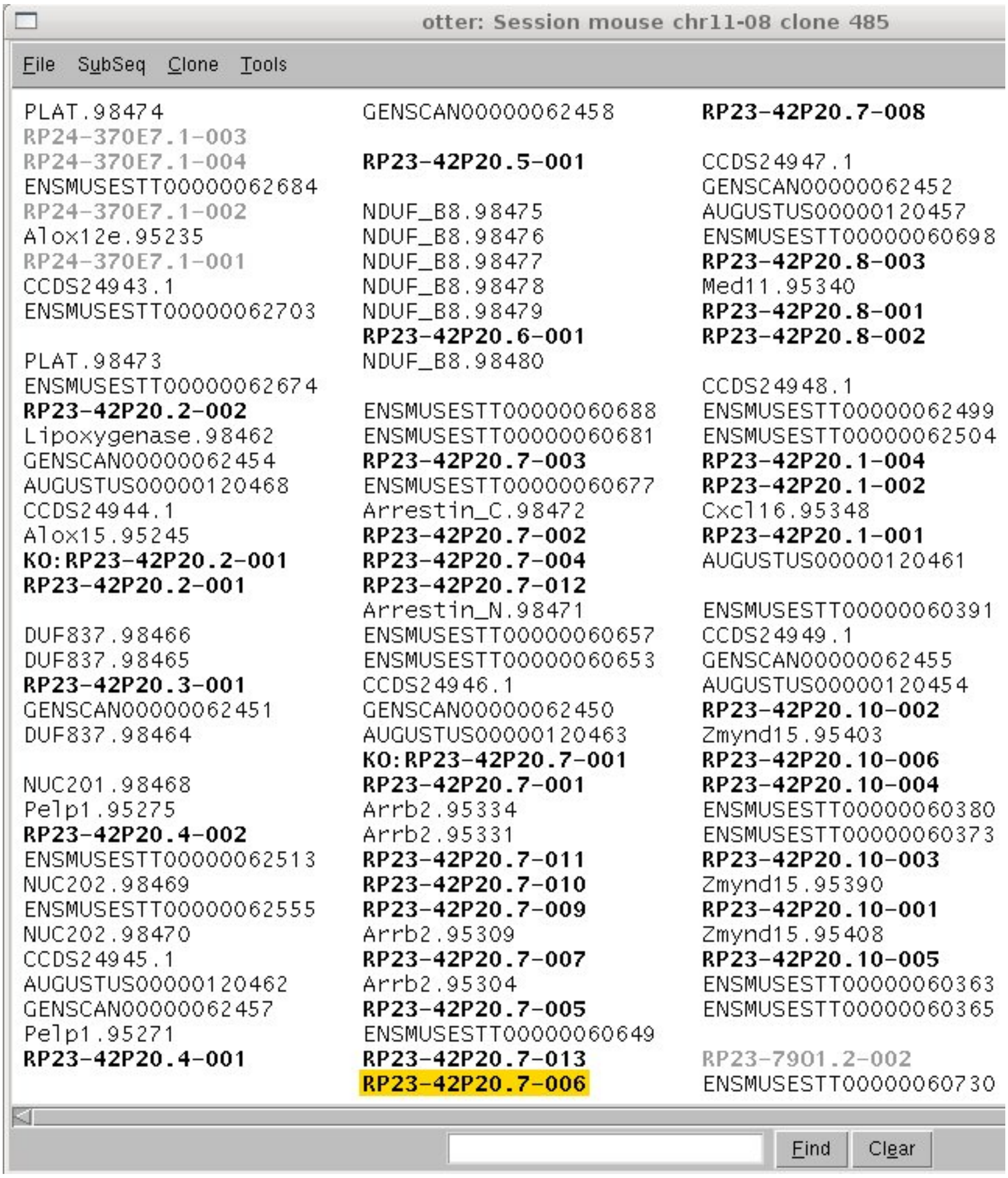
Alias(es): bM42P20.7 RP23-42P20.7

Remarks:

```
Protein   Sw:P32120
   cDNA   Em:AK097542.1
```

```
>RP23-42P20.7-006
MGEKPGTRVFKKSSPNCKLTVYLGKRDFVDHLDKVDPVVFVTLTCAFRYGREDLDVLGLS
FRKDLFIATYQAFPPMPNPPRPPTRLQDRLLKKLGQHAHPFFFTIPQNLPCSVTLQPGPE
DTGKACGVDFEIRAFCAKSIEEKSHKRNSVRLIIRKVQFAPETPGPQPSAETTRHFLMSD
RRSLHLEASLDKELYYHGEPLNVNVHVTNNSAKTVKKIRVSVRQYADICLFSTAQYKCPV
AQLEQDDQVSPSSTFCKVYTITPLLSDNREKRGLALDGQLKHEDTNLASSTIVKEGANKE
VLGILVSYRVKVKLVVSRGGDVSVELPFVLMHPKPHDHITLPRPQSAPIHPPLLCPSAPR
ETDVPVDTNLIEFDTNYATDDDIVFEDFARLRLKGMKDDDCDDQFC*
```

The main Otterlace window shows all the transcript models in the genomic region that is viewed. Those in bold are manual annotations.



PLAT.98474
RP24-370E7.1-003
RP24-370E7.1-004
ENSMUSESTT00000062684
RP24-370E7.1-002
Alox12e.95235
RP24-370E7.1-001
CCDS24943.1
ENSMUSESTT00000062703

PLAT.98473
ENSMUSESTT00000062674
RP23-42P20.2-002
Lipoxygenase.98462
GENSCAN00000062454
AUGUSTUS00000120468
CCDS24944.1
Alox15.95245
KO:RP23-42P20.2-001
RP23-42P20.2-001

DUF837.98466
DUF837.98465
RP23-42P20.3-001
GENSCAN00000062451
DUF837.98464

NUC201.98468
Pelp1.95275
RP23-42P20.4-002
ENSMUSESTT00000062513
NUC202.98469
ENSMUSESTT00000062555
NUC202.98470
CCDS24945.1
AUGUSTUS00000120462
GENSCAN00000062457
Pelp1.95271
RP23-42P20.4-001

GENSCAN00000062458

RP23-42P20.5-001

NDUF_B8.98475
NDUF_B8.98476
NDUF_B8.98477
NDUF_B8.98478
NDUF_B8.98479
RP23-42P20.6-001
NDUF_B8.98480

ENSMUSESTT00000060688
ENSMUSESTT00000060681
RP23-42P20.7-003
ENSMUSESTT00000060677
Arrestin_C.98472
RP23-42P20.7-002
RP23-42P20.7-004
RP23-42P20.7-012
Arrestin_N.98471
ENSMUSESTT00000060657
ENSMUSESTT00000060653
CCDS24946.1
GENSCAN00000062450
AUGUSTUS00000120463
KO:RP23-42P20.7-001
RP23-42P20.7-001
Arrb2.95334
Arrb2.95331
RP23-42P20.7-011
RP23-42P20.7-010
RP23-42P20.7-009
Arrb2.95309
RP23-42P20.7-007
Arrb2.95304
RP23-42P20.7-005
ENSMUSESTT00000060649
RP23-42P20.7-013
RP23-42P20.7-006

RP23-42P20.7-008

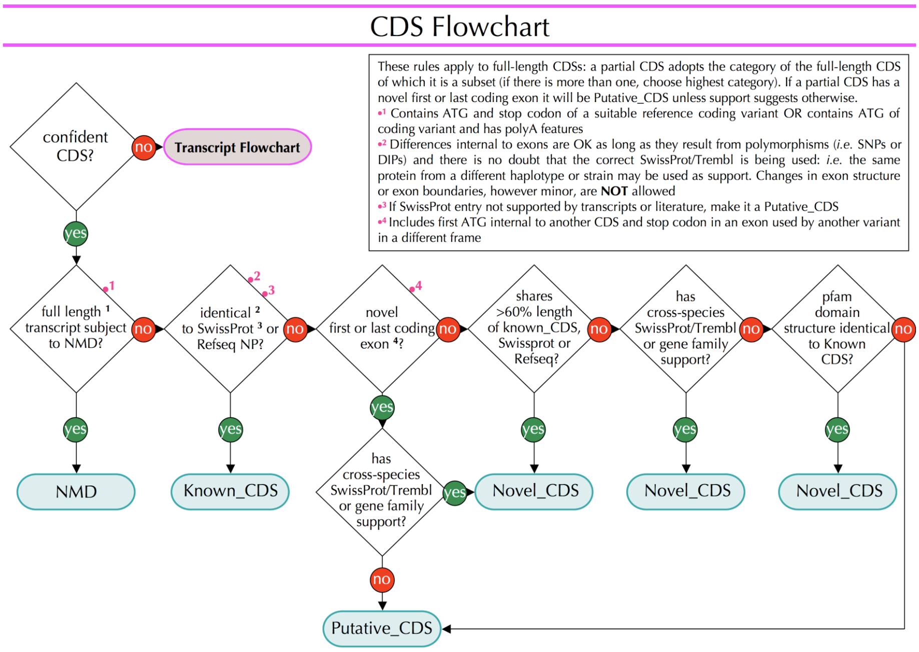
CCDS24947.1
GENSCAN00000062452
AUGUSTUS00000120457
ENSMUSESTT00000060698
RP23-42P20.8-003
Med11.95340
RP23-42P20.8-001
RP23-42P20.8-002

CCDS24948.1
ENSMUSESTT00000062499
ENSMUSESTT00000062504
RP23-42P20.1-004
RP23-42P20.1-002
Cxcl16.95348
RP23-42P20.1-001
AUGUSTUS00000120461

ENSMUSESTT00000060391
CCDS24949.1
GENSCAN00000062455
AUGUSTUS00000120454
RP23-42P20.10-002
Zmynd15.95403
RP23-42P20.10-006
RP23-42P20.10-004
ENSMUSESTT00000060380
ENSMUSESTT00000060373
RP23-42P20.10-003
Zmynd15.95390
RP23-42P20.10-001
Zmynd15.95408
RP23-42P20.10-005
ENSMUSESTT00000060363
ENSMUSESTT00000060365

RP23-7901.2-002
ENSMUSESTT00000060730

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



CDS Flowchart

These rules apply to full-length CDSs: a partial CDS adopts the category of the full-length CDS of which it is a subset (if there is more than one, choose highest category). If a partial CDS has a novel first or last coding exon it will be Putative_CDS unless support suggests otherwise.

•[1] Contains ATG and stop codon of a suitable reference coding variant OR contains ATG of coding variant and has polyA features

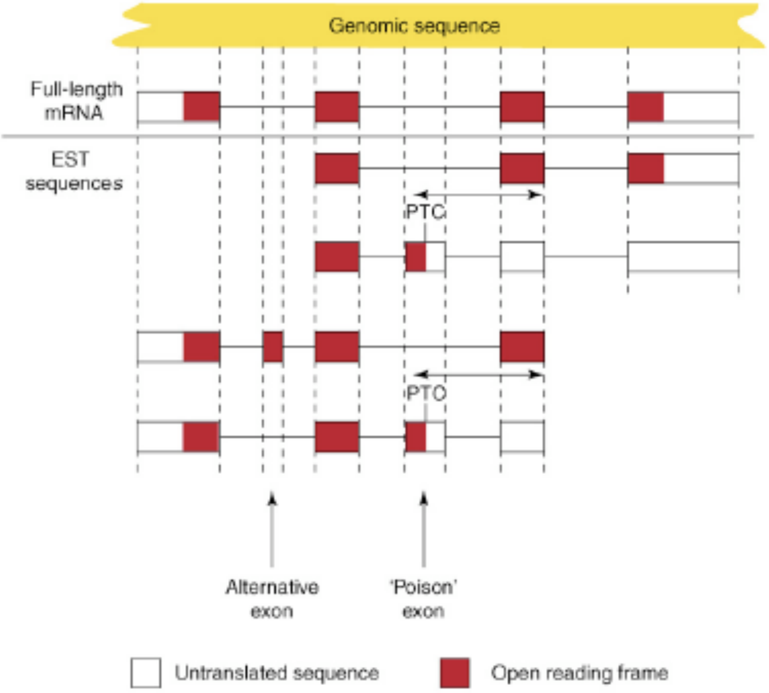
•[2] Differences internal to exons are OK as long as they result from polymorphisms (*i.e.* SNPs or DIPs) and there is no doubt that the correct SwissProt/Trembl is being used: *i.e.* the same protein from a different haplotype or strain may be used as support. Changes in exon structure or exon boundaries, however minor, are **NOT** allowed

•[3] If SwissProt entry not supported by transcripts or literature, make it a Putative_CDS

•[4] Includes first ATG internal to another CDS and stop codon in an exon used by another variant in a different frame

confident CDS?
no
Transcript Flowchart
yes
full length ¹ transcript subject to NMD?
•1
no
yes
NMD
identical ² to SwissProt ³ or Refseq NP?
•2
•3
no
yes
Known_CDS
novel first or last coding exon ⁴?
•4
no
yes
has cross-species SwissProt/Trembl or gene family support?
yes
no
Putative_CDS
shares >60% length of known_CDS, Swissprot or Refseq?
no
yes
Novel_CDS
has cross-species SwissProt/Trembl or gene family support?
no
yes
Novel_CDS
pfam domain structure identical to Known CDS?
no
yes
Novel_CDS

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

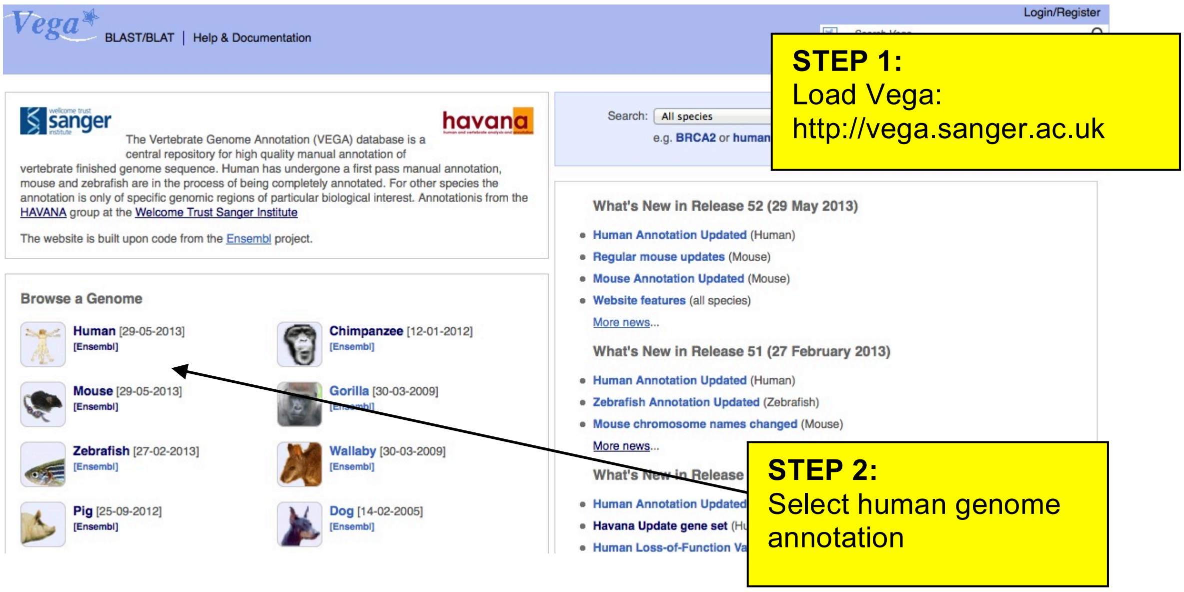


The exact mechanisms behind NMD have not been elucidated and so we retain the CDS in our gene models.

The Vertebrate Genome Annotation (VEGA) Database

Worked example:

1. View the CIZ1 locus. How many transcripts are there in Vega compared to Ensembl? Which transcript is the CCDS? Export this peptide sequence.


Vega
BLAST/BLAT | Help & Documentation
Login/Register
sanger
havana
Search: All species
e.g. BRCA2 or human
Browse a Genome
Human [29-05-2013] [Ensembl]
Chimpanzee [12-01-2012] [Ensembl]
Mouse [29-05-2013] [Ensembl]
Gorilla [30-03-2009] [Ensembl]
Zebrafish [27-02-2013] [Ensembl]
Wallaby [30-03-2009] [Ensembl]
Pig [25-09-2012] [Ensembl]
Dog [14-02-2005] [Ensembl]

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality manual annotation of vertebrate finished genome sequence. Human has undergone a first pass manual annotation, mouse and zebrafish are in the process of being completely annotated. For other species the annotation is only of specific genomic regions of particular biological interest. Annotationis from the HAVANA group at the Welcome Trust Sanger Institute

The website is built upon code from the Ensembl project.

What's New in Release 52 (29 May 2013)

- Human Annotation Updated (Human)
- Regular mouse updates (Mouse)
- Mouse Annotation Updated (Mouse)
- Website features (all species)

More news...

What's New in Release 51 (27 February 2013)

- Human Annotation Updated (Human)
- Zebrafish Annotation Updated (Zebrafish)
- Mouse chromosome names changed (Mouse)

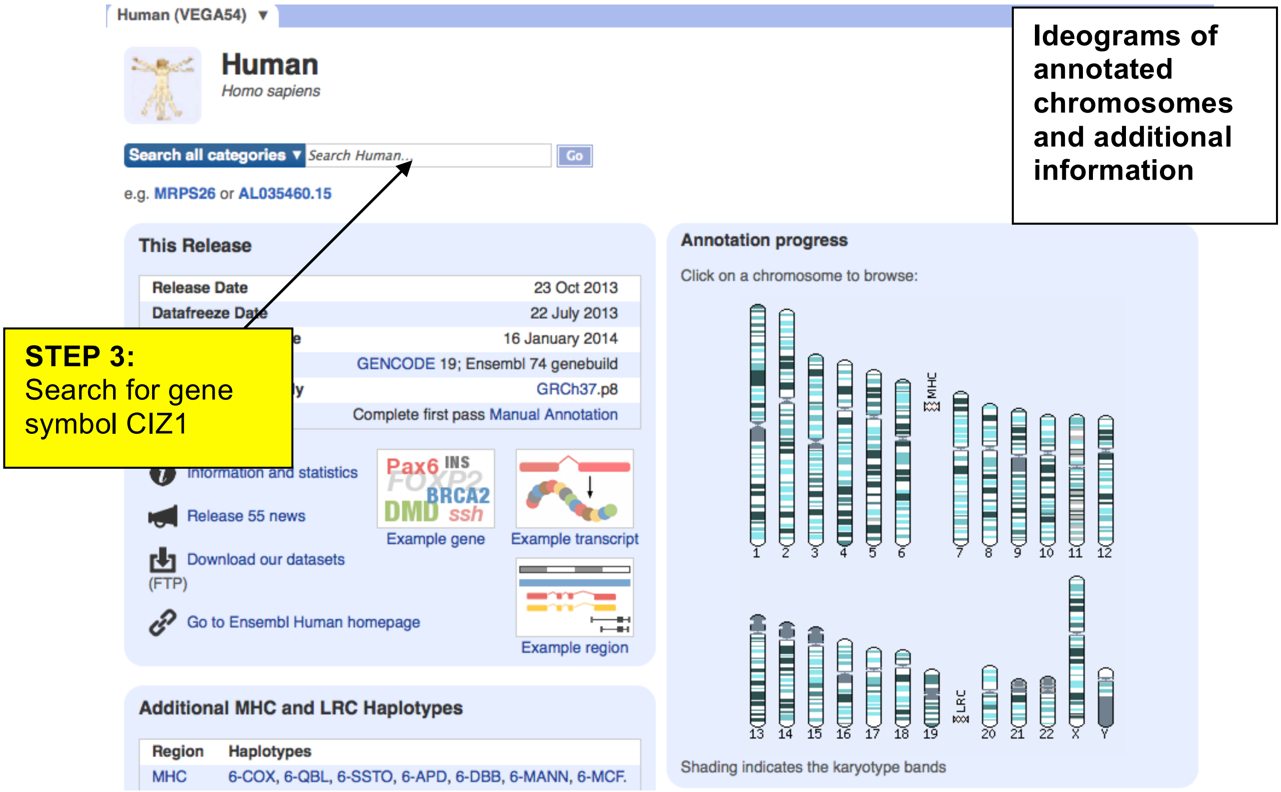
More news...

What's New in Release

- Human Annotation Updated
- Havana Update gene set (Hu
- Human Loss-of-Function Va

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

STEP 2:
Select human genome annotation


Human (VEGA54)
Human
Homo sapiens
Search all categories
Search Human...
Go
Information and statistics
Release 55 news
Download our datasets (FTP)
Go to Ensembl Human homepage
Pax6 INS FOXP2 BRCA2 DMD ssh
Example gene
Example transcript
Example region
Click on a chromosome to browse:
1
2
3
4
5
6
MHC
7
8
9
10
11
12
13
14
15
16
17
18
19
LRC
20
21
22
X
Y
Shading indicates the karyotype bands

e.g. MRPS26 or AL035460.15

This Release

Release Date	23 Oct 2013
Datafreeze Date	22 July 2013
e	16 January 2014
	GENCODE 19; Ensembl 74 genebuild
y	GRCh37.p8
	Complete first pass Manual Annotation

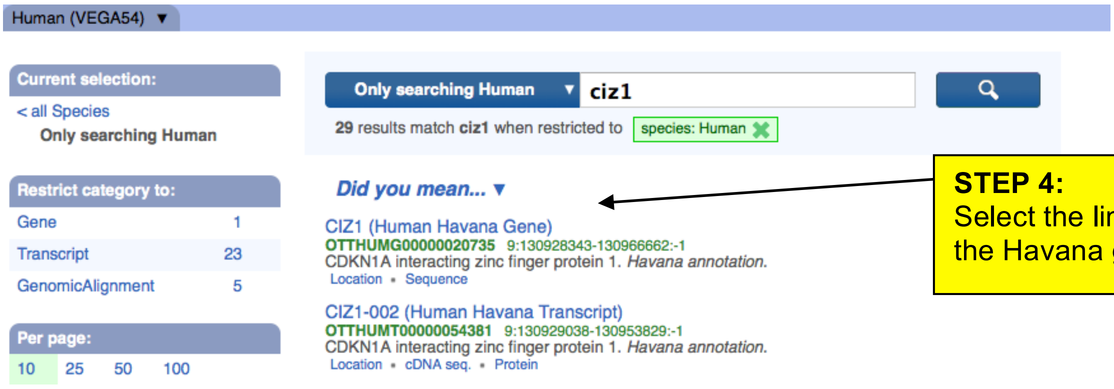
Annotation progress

Additional MHC and LRC Haplotypes

Region	Haplotypes
MHC	6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF.

Ideograms of annotated chromosomes and additional information

STEP 3:
Search for gene symbol CIZ1



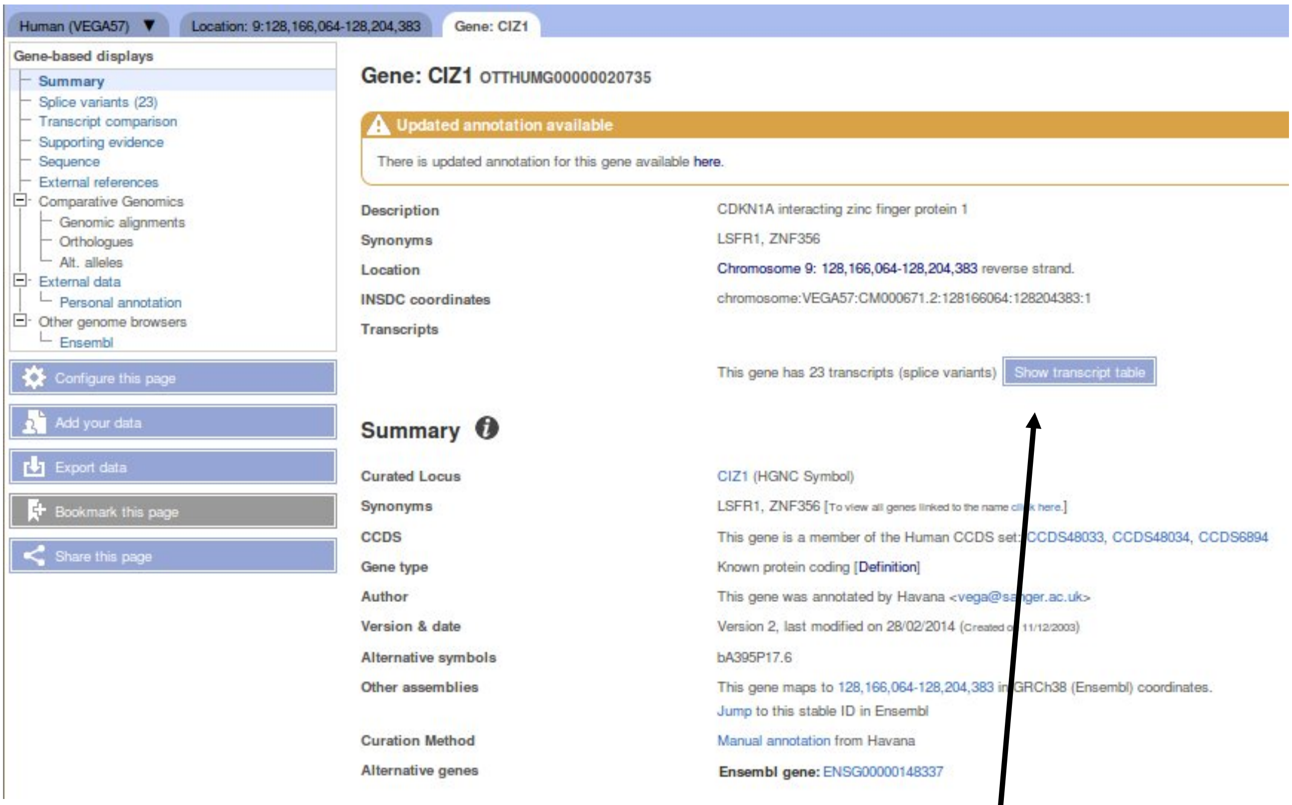
29 results match **ciz1** when restricted to species: Human ✖

Did you mean... ▼

CIZ1 (Human Havana Gene)
OTTHUMG00000020735 9:130928343-130966662:-1
CDKN1A interacting zinc finger protein 1. *Havana annotation.*
Location • Sequence

CIZ1-002 (Human Havana Transcript)
OTTHUMT00000054381 9:130929038-130953829:-1
CDKN1A interacting zinc finger protein 1. *Havana annotation.*
Location • cDNA seq. • Protein

STEP 4:
Select the link to the Havana gene



Gene: CIZ1 OTTHUMG00000020735

Updated annotation available
There is updated annotation for this gene available here.

Description	CDKN1A interacting zinc finger protein 1
Synonyms	LSFR1, ZNF356
Location	Chromosome 9: 128,166,064-128,204,383 reverse strand.
INSDC coordinates	chromosome:VEGA57:CM000671.2:128166064:128204383:1
Transcripts	
	This gene has 23 transcripts (splice variants) Show transcript table

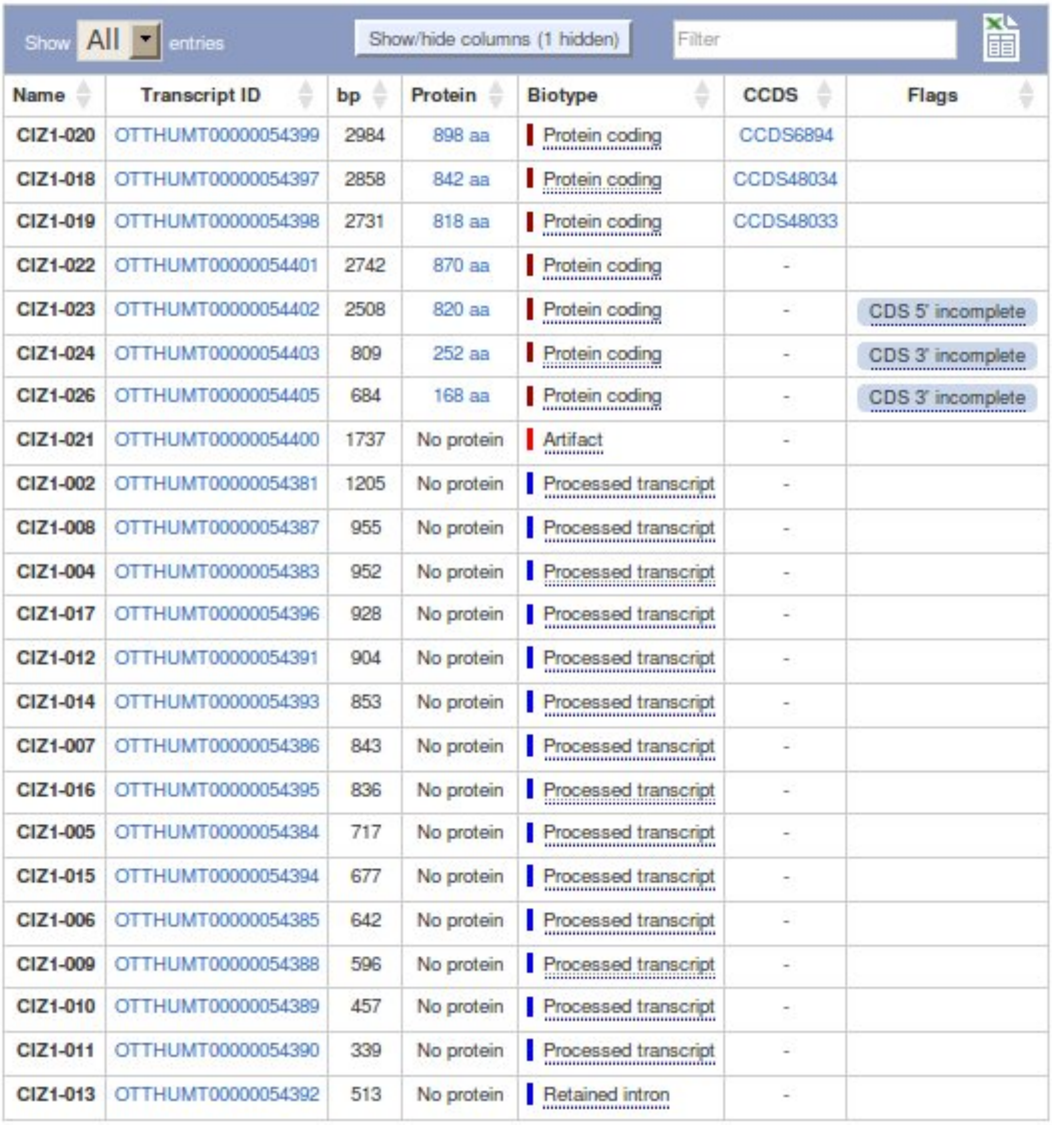
Summary

Curated Locus	CIZ1 (HGNC Symbol)
Synonyms	LSFR1, ZNF356 [To view all genes linked to the name click here.]
CCDS	This gene is a member of the Human CCDS set: CCDS48033, CCDS48034, CCDS6894
Gene type	Known protein coding [Definition]
Author	This gene was annotated by Havana <vega@sanger.ac.uk>
Version & date	Version 2, last modified on 28/02/2014 (Created on 11/12/2003)
Alternative symbols	bA395P17.6
Other assemblies	This gene maps to 128,166,064-128,204,383 in GRCh38 (Ensembl) coordinates. Jump to this stable ID in Ensembl
Curation Method	Manual annotation from Havana
Alternative genes	**Ensembl gene:** ENSG00000148337

STEP 5:
The page opens at the gene summary. Click on the Show transcript Table link to display all the manually annotated transcripts.

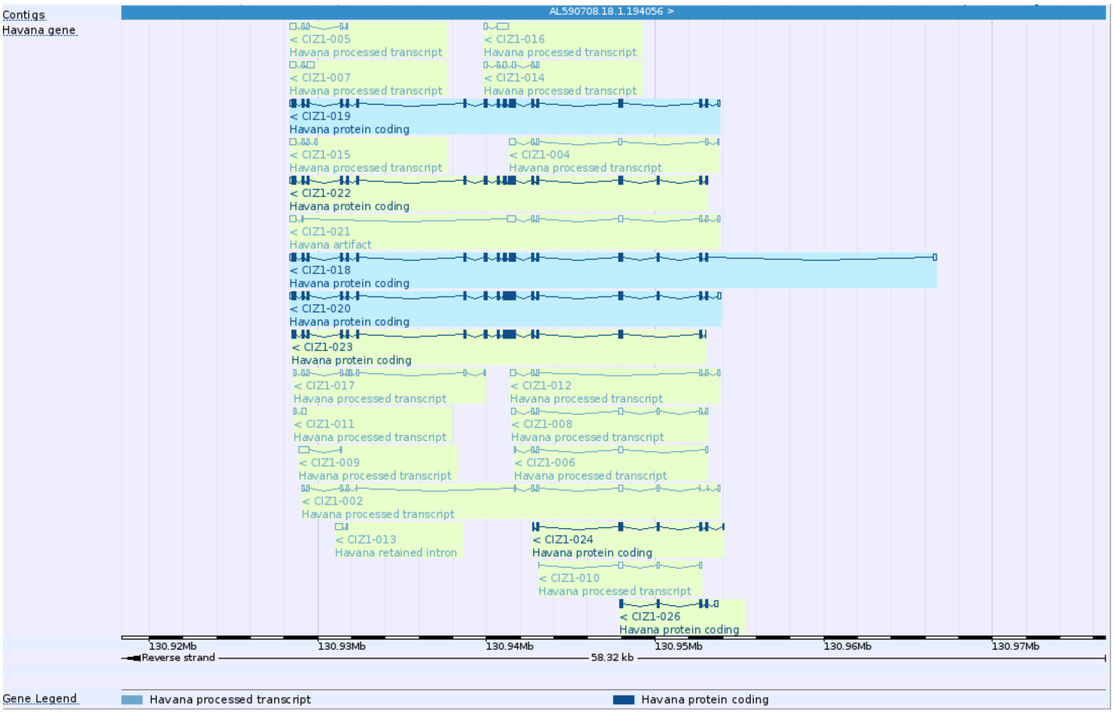
List of manually curated transcripts and their corresponding transcript models.

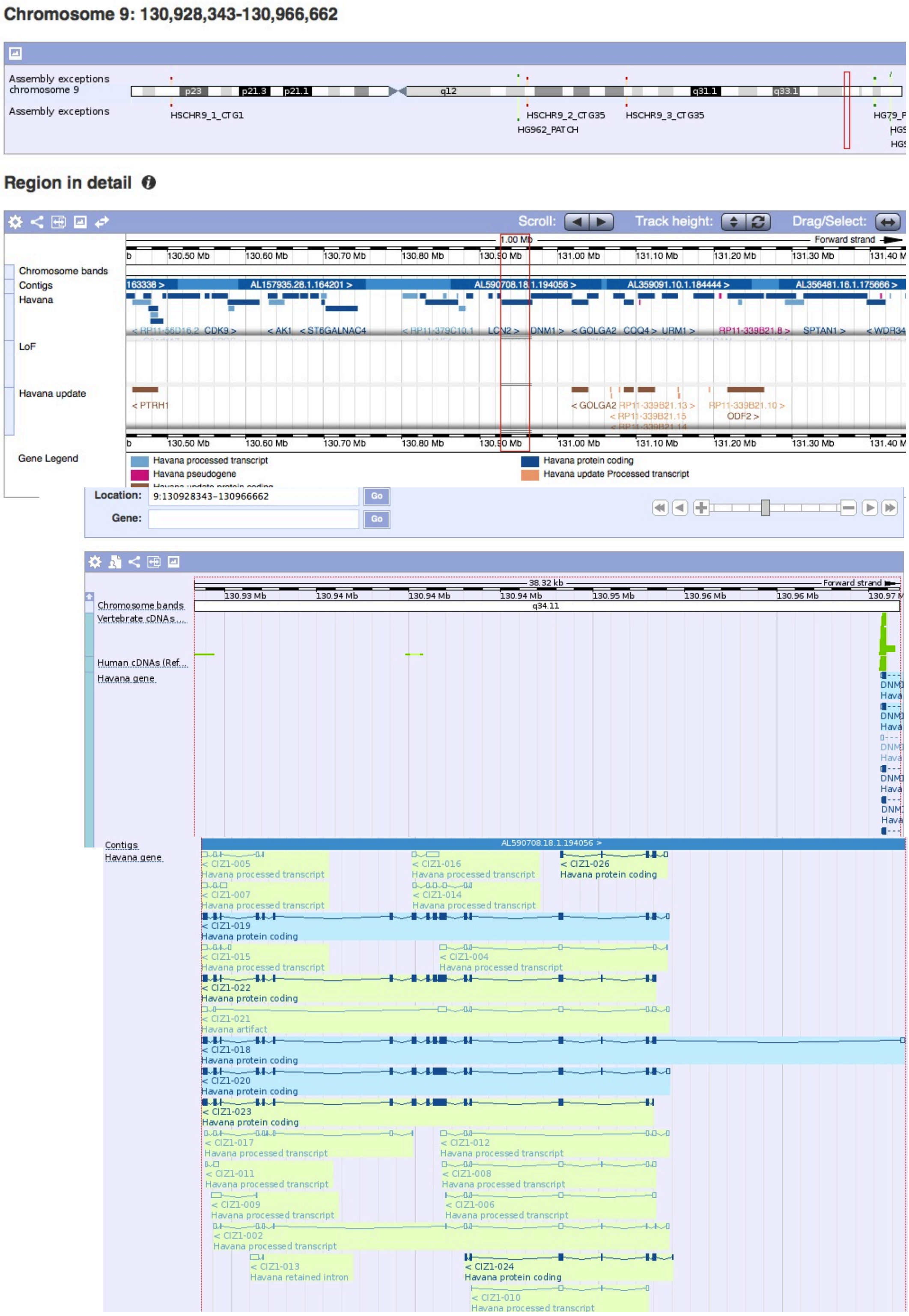
Scroll to the bottom of the page to see the gene summary on the genome.



Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
CIZ1-020	OTTHUMT00000054399	2984	898 aa	Protein coding	CCDS6894	
CIZ1-018	OTTHUMT00000054397	2858	842 aa	Protein coding	CCDS48034	
CIZ1-019	OTTHUMT00000054398	2731	818 aa	Protein coding	CCDS48033	
CIZ1-022	OTTHUMT00000054401	2742	870 aa	Protein coding	-	
CIZ1-023	OTTHUMT00000054402	2508	820 aa	Protein coding	-	CDS 5' incomplete
CIZ1-024	OTTHUMT00000054403	809	252 aa	Protein coding	-	CDS 3' incomplete
CIZ1-026	OTTHUMT00000054405	684	168 aa	Protein coding	-	CDS 3' incomplete
CIZ1-021	OTTHUMT00000054400	1737	No protein	Artifact	-	
CIZ1-002	OTTHUMT00000054381	1205	No protein	Processed transcript	-	
CIZ1-008	OTTHUMT00000054387	955	No protein	Processed transcript	-	
CIZ1-004	OTTHUMT00000054383	952	No protein	Processed transcript	-	
CIZ1-017	OTTHUMT00000054396	928	No protein	Processed transcript	-	
CIZ1-012	OTTHUMT00000054391	904	No protein	Processed transcript	-	
CIZ1-014	OTTHUMT00000054393	853	No protein	Processed transcript	-	
CIZ1-007	OTTHUMT00000054386	843	No protein	Processed transcript	-	
CIZ1-016	OTTHUMT00000054395	836	No protein	Processed transcript	-	
CIZ1-005	OTTHUMT00000054384	717	No protein	Processed transcript	-	
CIZ1-015	OTTHUMT00000054394	677	No protein	Processed transcript	-	
CIZ1-006	OTTHUMT00000054385	642	No protein	Processed transcript	-	
CIZ1-009	OTTHUMT00000054388	596	No protein	Processed transcript	-	
CIZ1-010	OTTHUMT00000054389	457	No protein	Processed transcript	-	
CIZ1-011	OTTHUMT00000054390	339	No protein	Processed transcript	-	
CIZ1-013	OTTHUMT00000054392	513	No protein	Retained intron	-	

CCDS member highlighted in blue





Chromosome 9: 130,928,343-130,966,662

Region in detail

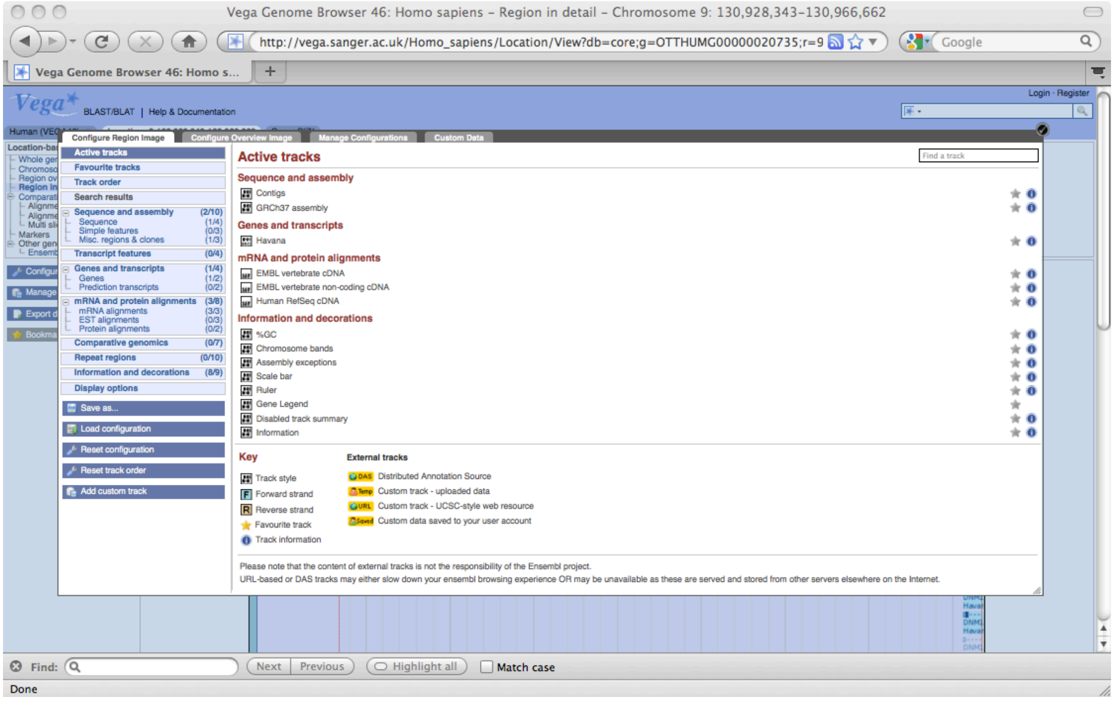
Location: 9:130928343-130966662 Go

Gene: Go

Vega shows 23 variants for CIZ1, 7 of which are protein coding (as shown by the solid blue boxes).

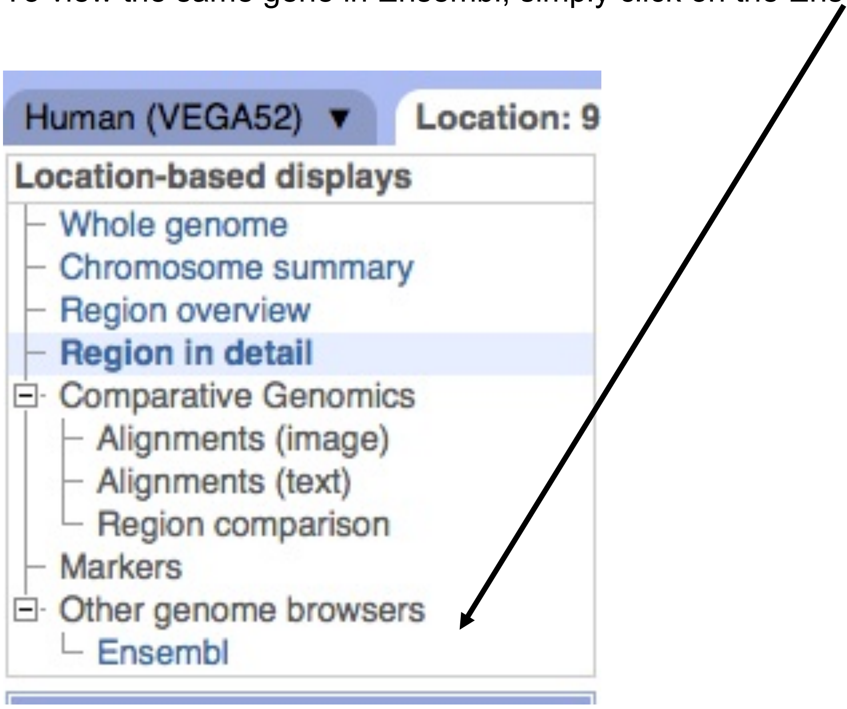
Go back up to the top of this page and click on the location link or tab. This will bring you through to the region in detail page.

Tracks may be switched on and off in the configure page found in the left-hand menu.

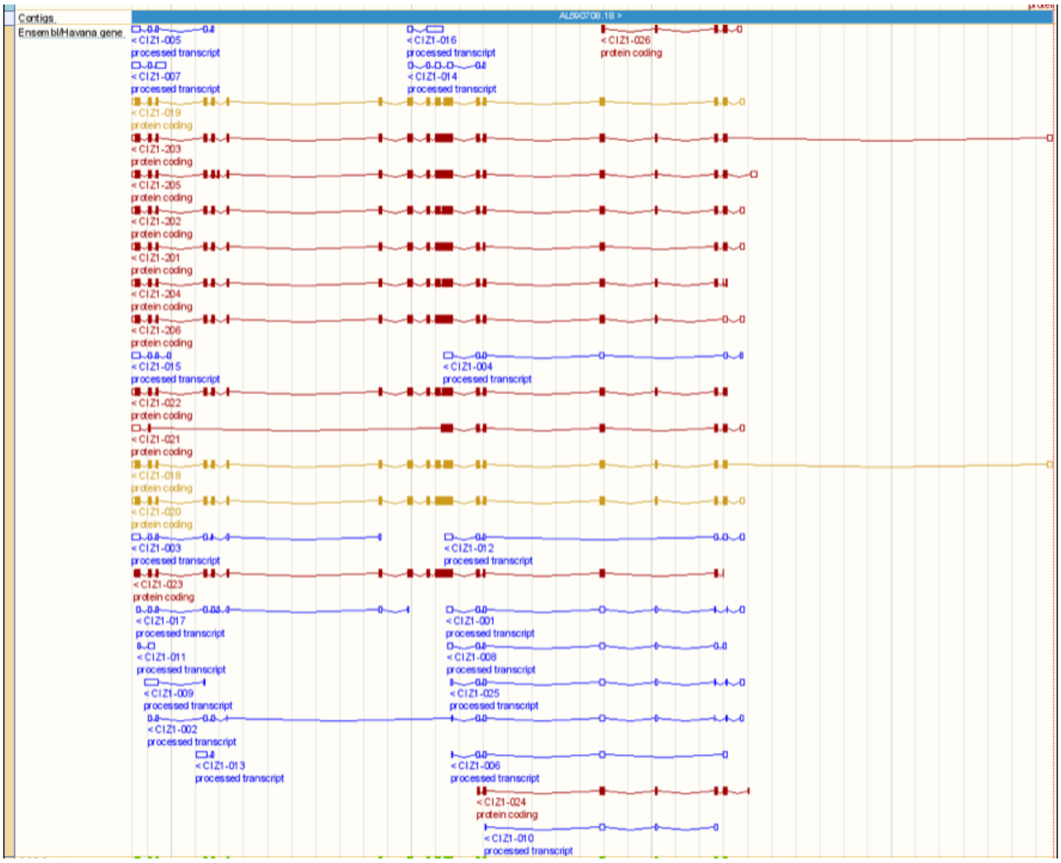


Please note that the content of external tracks is not the responsibility of the Ensembl project.
URL-based or DAS tracks may either slow down your ensembl browsing experience OR may be unavailable as these are served and stored from other servers elsewhere on the Internet.

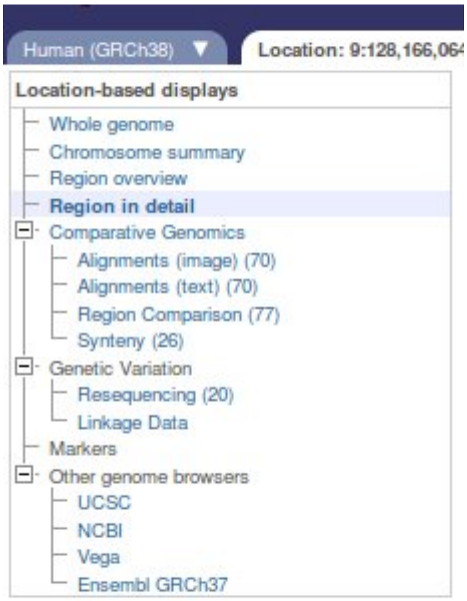
To view the same gene in Ensembl, simply click on the Ensembl link in the left hand menu.



Tracks in Ensembl are shown in red, gold and blue. Ensembl has 8 predictions in red (Protein Coding), in gold 3 Protein Coding Ensembl/Havana merges (agrees with a Vega coding transcript), and 15 processed transcript gene variants from the Havana manual annotation.
Vega tracks may also be displayed in Ensembl as a separate track, as may the CCDS gene set.

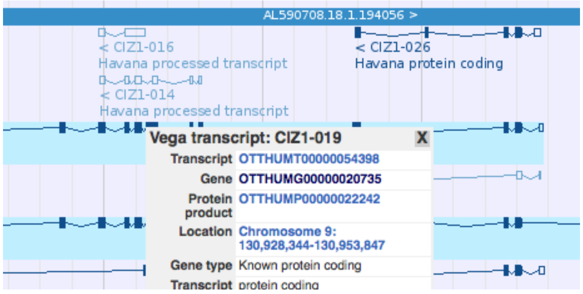

Contigs
Ensembl/Havana gene
AL590708.18 >
< CIZ1-005 processed transcript
< CIZ1-007 processed transcript
< CIZ1-019 protein coding
< CIZ1-203 protein coding
< CIZ1-205 protein coding
< CIZ1-202 protein coding
< CIZ1-201 protein coding
< CIZ1-204 protein coding
< CIZ1-206 protein coding
< CIZ1-015 processed transcript
< CIZ1-022 protein coding
< CIZ1-021 protein coding
< CIZ1-018 protein coding
< CIZ1-020 protein coding
< CIZ1-003 processed transcript
< CIZ1-023 protein coding
< CIZ1-017 processed transcript
< CIZ1-011 processed transcript
< CIZ1-009 processed transcript
< CIZ1-002 processed transcript
< CIZ1-013 processed transcript
< CIZ1-016 processed transcript
< CIZ1-014 processed transcript
< CIZ1-026 protein coding
< CIZ1-004 processed transcript
< CIZ1-012 processed transcript
< CIZ1-001 processed transcript
< CIZ1-008 processed transcript
< CIZ1-025 processed transcript
< CIZ1-006 processed transcript
< CIZ1-024 protein coding
< CIZ1-010 processed transcript

You may export data easily from Vega.



STEP 6:
Click back to **Vega** by clicking on the Vega link at the side and jump to Vega.

For example export the protein sequence for CIZ1 main variant.

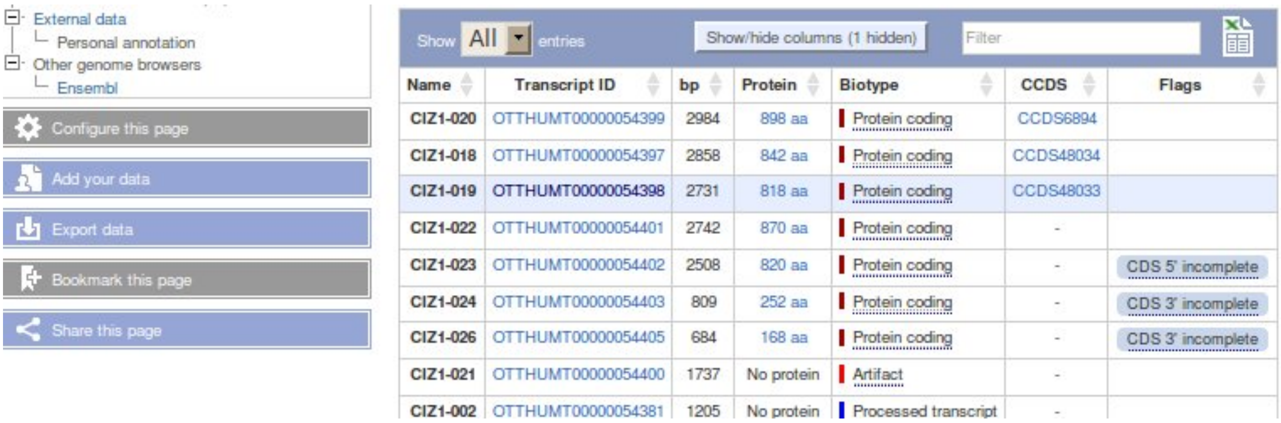


Vega transcript: CIZ1-019 X

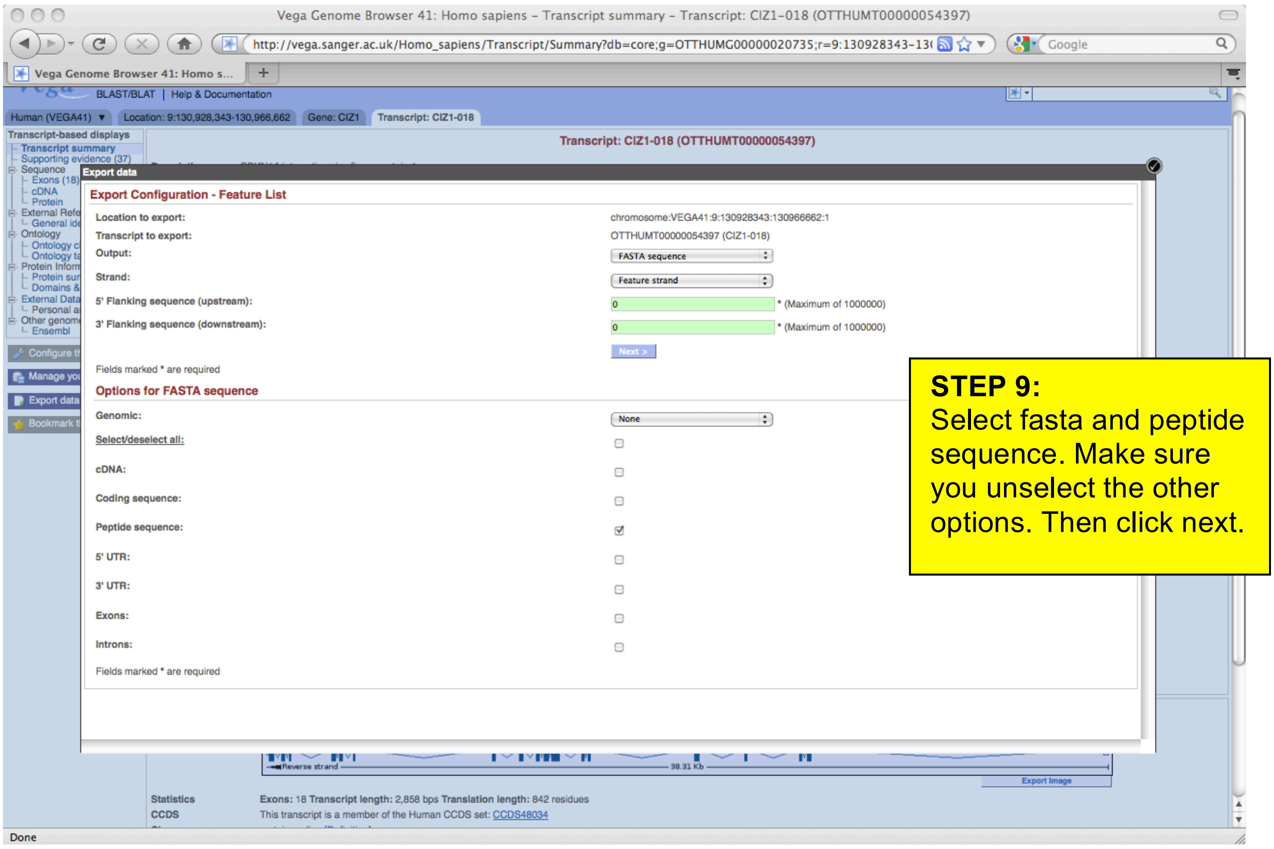
Transcript	OTTHUMT00000054398
Gene	OTTHUMG00000020735
Protein product	OTTHUMP00000022242
Location	Chromosome 9: 130,928,344-130,953,847
Gene type	Known protein coding
Transcript	protein coding

STEP 7: Click onto any of the gene objects and select the gene link.

STEP 8: Click on any protein coding transcript to bring up the protein summary and then click on the left-hand export data link.



Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
CIZ1-020	OTTHUMT00000054399	2984	898 aa	Protein coding	CCDS6894	
CIZ1-018	OTTHUMT00000054397	2858	842 aa	Protein coding	CCDS48034	
CIZ1-019	OTTHUMT00000054398	2731	818 aa	Protein coding	CCDS48033	
CIZ1-022	OTTHUMT00000054401	2742	870 aa	Protein coding	-	
CIZ1-023	OTTHUMT00000054402	2508	820 aa	Protein coding	-	CDS 5' incomplete
CIZ1-024	OTTHUMT00000054403	809	252 aa	Protein coding	-	CDS 3' incomplete
CIZ1-026	OTTHUMT00000054405	684	168 aa	Protein coding	-	CDS 3' incomplete
CIZ1-021	OTTHUMT00000054400	1737	No protein	Artifact	-	
CIZ1-002	OTTHUMT00000054381	1205	No protein	Processed transcript	-	



Vega Genome Browser 41: Homo sapiens – Transcript summary – Transcript: CIZ1-018 (OTTHUMT00000054397)

http://vega.sanger.ac.uk/Homo_sapiens/Transcript/Summary?db=core;g=OTTHUMG00000020735;r=9:130928343-13(

Human (VEGA41) | Location: 9:130,928,343-130,966,662 | Gene: CIZ1 | Transcript: CIZ1-018

Transcript: CIZ1-018 (OTTHUMT00000054397)

Export data

Export Configuration - Feature List

Location to export:	chromosome:VEGA41:9:130928343:130966662:1
Transcript to export:	OTTHUMT00000054397 (CIZ1-018)
Output:	FASTA sequence
Strand:	Feature strand
5' Flanking sequence (upstream):	0 * (Maximum of 1000000)
3' Flanking sequence (downstream):	0 * (Maximum of 1000000)

Fields marked * are required

Options for FASTA sequence

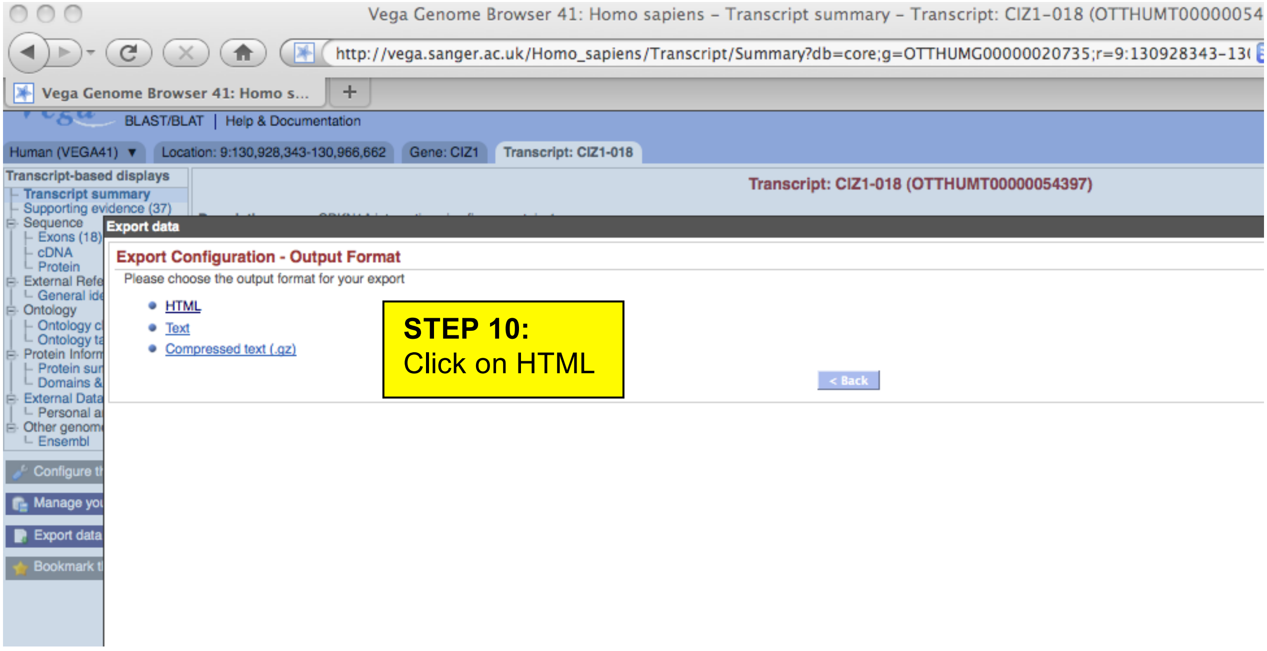
Genomic:	None
Select/deselect all:	□
cDNA:	□
Coding sequence:	□
Peptide sequence:	☑
5' UTR:	□
3' UTR:	□
Exons:	□
Introns:	□

Fields marked * are required

Statistics Exons: 18 Transcript length: 2,858 bps Translation length: 842 residues

CCDS This transcript is a member of the Human CCDS set: CCDS48034

STEP 9:
Select fasta and peptide sequence. Make sure you unselect the other options. Then click next.



Vega Genome Browser 41: Homo sapiens – Transcript summary – Transcript: CIZ1-018 (OTTHUMT00000054

http://vega.sanger.ac.uk/Homo_sapiens/Transcript/Summary?db=core;g=OTTHUMG00000020735;r=9:130928343-13(

Human (VEGA41) | Location: 9:130,928,343-130,966,662 | Gene: CIZ1 | Transcript: CIZ1-018

Transcript: CIZ1-018 (OTTHUMT00000054397)

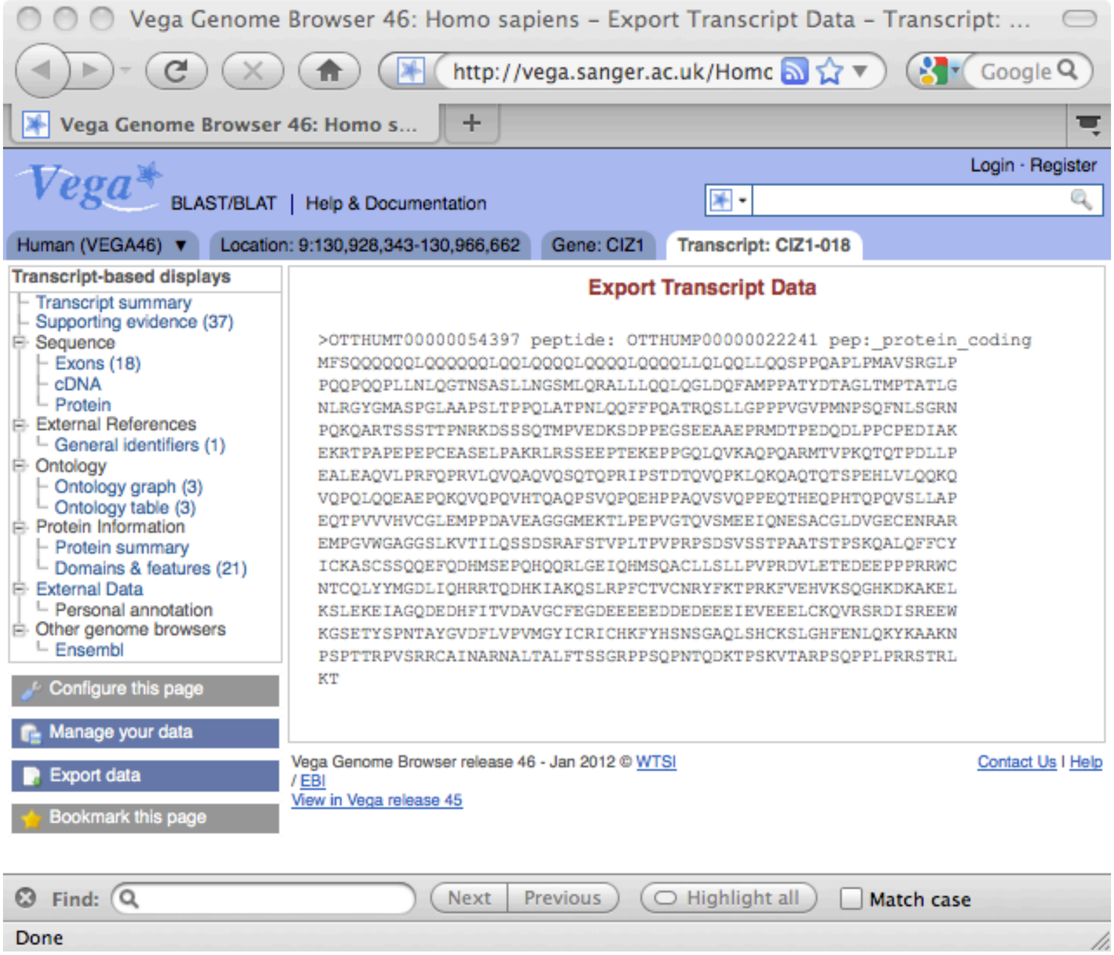
Export data

Export Configuration - Output Format

Please choose the output format for your export

- HTML
- Text
- Compressed text (.gz)

STEP 10:
Click on HTML



Export Transcript Data

```
>OTTHUMT00000054397 peptide: OTTHUMP00000022241 pep:_protein_coding
MFSQQQQQQLQQQQQQLQQLQQQQLQQQQLQQQQLLQLQQLLQQSPPQAPLPMAVSRGLP
PQQPQQPLLNLQGTNSASLLNGSMLQRALLLQQLQGLDQFAMPPATYDTAGLTMPTATLG
NLRGYGMASPGLAAPSLTPPQLATPNLQQFFPQATRQSLLGPPPVGVPMNPSQFNLSGRN
PQKQARTSSSTTPNRKDSSSQTMPVEDKSDPPEGSEEAAEPRMDTPEDQDLPPCPEDIAK
EKRTPAPEPEPCEASELPAKRLRSSEEPTEKEPPGQLQVKAQPQARMTVPKQTQTPDLLP
EALEAQVLPRFQPRVLQVQAQVQSQTQPRIPSTDTQVQPKLQKQAQTQTSPEHLVLQQKQ
VQPQLQQEAEPQKQVQPQVHTQAQPSVQPQEHPPAQVSVQPPEQTHEQPHTQPQVSLLAP
EQTPVVVHVCGLEMPPDAVEAGGGMEKTLPEPVGTQVSMEEIQNESACGLDVGECENRAR
EMPGVWGAGGSLKVTILQSSDSRAFSTVPLTPVPRPSDSVSSTPAATSTPSKQALQFFCY
ICKASCSSQQEFQDHMSEPQHQQRLGEIQHMSQACLLSLLPVPRDVLETEDEEPPPRRWC
NTCQLYYMGDLIQHRRTQDHKIAKQSLRPFCTVCNRYFKTPRKFVEHVKSQGHKDKAKEL
KSLEKEIAGQDEDHFITVDAVGCFEGDEEEEEDDEDEEEIEVEEELCKQVRSRDISREEW
KGSETYSPNTAYGVDFLVPVMGYICRICHKFYHSNSGAQLSHCKSLGHFENLQKYKAAKN
PSPTTRPVSRRCAINARNALTALFTSSGRPPSQPNTQDKTPSKVTARPSQPPLPRRSTRL
KT
```

Vega Genome Browser release 46 - Jan 2012 © WTSI / EBI
View in Vega release 45

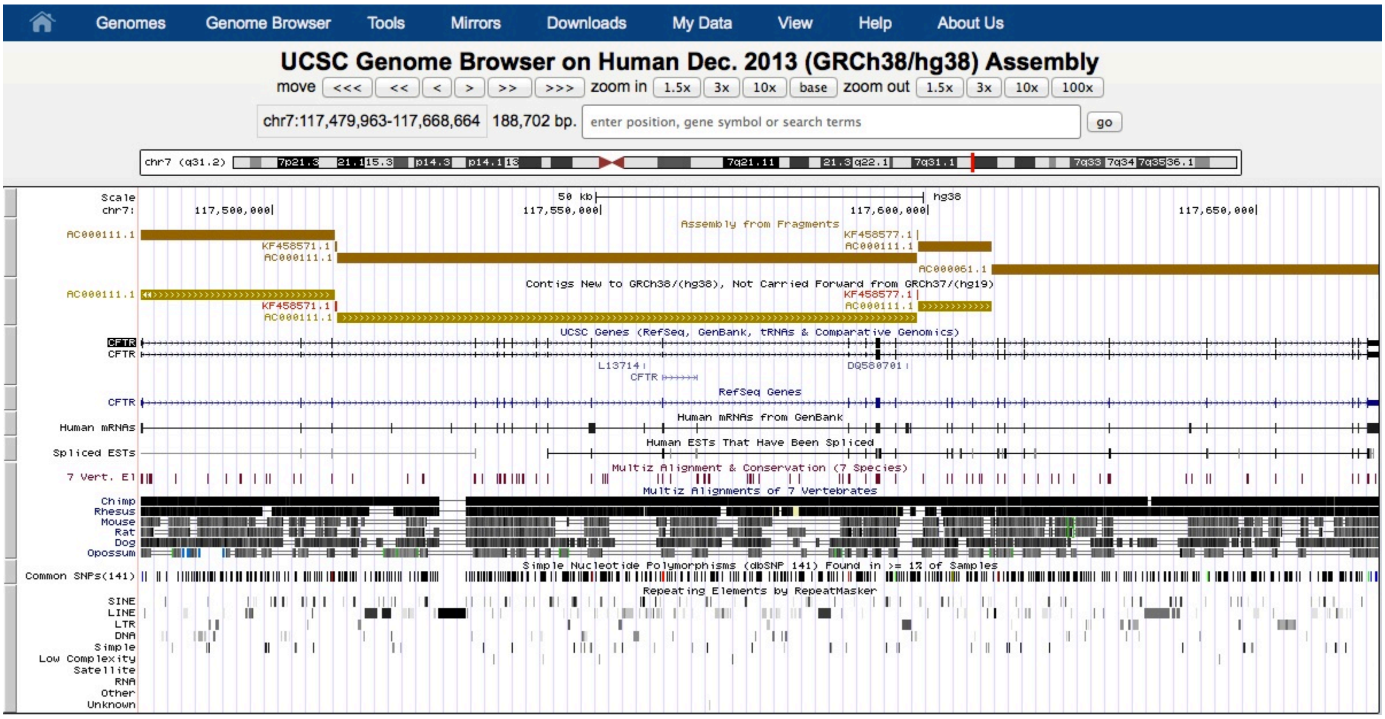
Contact Us | Help

Sequence may be copied and pasted into other programs for further analysis.

UCSC Human Genome Browser

The University of California Santa Cruz (UCSC) Genome Browser at **http://genome.ucsc.edu** is a web-based set of tools providing access to a database of genome sequence and annotations for visualization, comparison and analysis by the scientific, medical and academic communities. There is also an official European mirror site: http://genome-euro.ucsc.edu. The primary mission of the site is to provide timely and convenient open access to high-quality human genome sequence and annotations in a framework that enables easy exploration from genome-wide down to the base level. Annotation datasets, or 'tracks', on the human genome cover conservation and evolutionary comparisons, gene models, regulation, expression, epigenetics and tissue differentiation, variation, phenotype and disease associations. A substantial contributor to UCSC has been participation in the ENCODE project as the designated data repository in the ENCODE Pilot (2003-2007) and as the Data Coordination Center (DCC) in the ENCODE whole-genome data production phase (2007-2012). All production ENCODE data is routed to UCSC for validation, quality review, database storage, visualization, and dissemination to other public databases. At this time more than 2700 distinct ENCODE experiments have been processed by the DCC and made publicly available.

Other organisms represented at the site include 13 primates, 33 other mammals, 17 non-mammalian vertebrates, 13 insects, 6 worms and 5 other invertebrates. There is also an Ebola virus browser built from viral strains from previous outbreaks as well as the 2014 outbreak. The UCSC browser also contains Neandertal sequence data and alignments to the human genome. The browser hosts mapping and sequence annotation tracks that describe assembly, gap and GC content for all organisms in the browser database. Additionally, for most organisms alignments are shown from RefSeq genes, mRNAs and ESTs from GenBank, and other gene or gene prediction tracks such as Ensembl Genes. For human and mouse assemblies, there is also a locally generated UCSC Genes track based upon RefSeq, GenBank and CCDS data. About half of the genomes hosted at UCSC include a multiple-sequence alignment track and pairwise genomic alignments between assemblies to further comparative and evolutionary investigations. Expression, regulation, variation and phenotype tracks are available for many of the assemblies. User data can be uploaded and visualized, and offer a data-hub mechanism allowing visualization of user data hosted remotely.



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr7:117,479,963-117,668,664 188,702 bp. enter position, gene symbol or search terms go

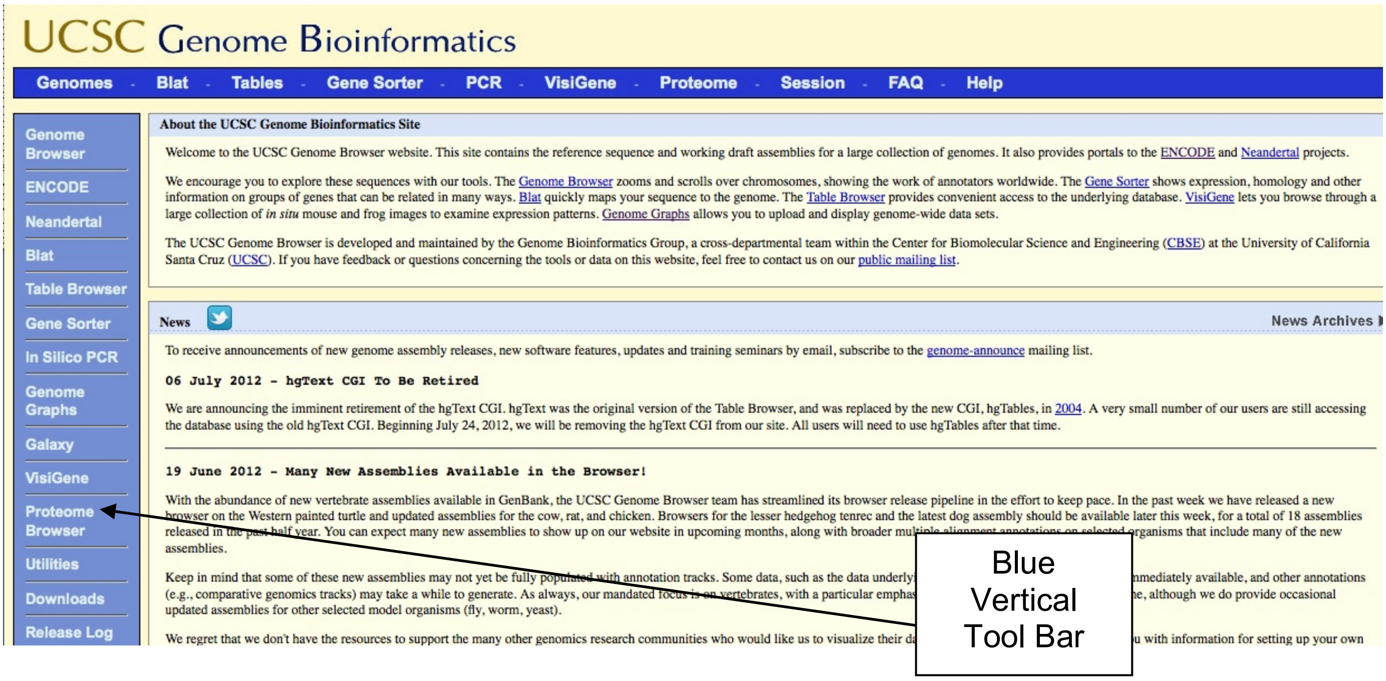
- Straight-forward feature display, easy to navigate
- Wide range of annotations (called “tracks”) including those supplied by other groups
- Good cross-species and evolutionary conservation annotations
- Expression data (GNF Atlas 2), microarray chip probe locations
- “Wiggle” tracks for continuous valued data
- All data available in bulk downloads or through Table Browser
- Ability to view own data
- Fast sequence searches using BLAT including paired sequences (isPCR)
- In situ images with transcription information (VisiGene Browser)
- Graphing of data on karyotype, like association and linkage test data
- Data hub for ENCODE data

While browsers can be very useful tools, they do not provide the definitive answers to every question and are not guaranteed to contain accurate data!

Practical Example

STEP 2:
Select “Genomes” from blue horizontal navigation bar (far left) and select human, Feb. 2009 GRCh37/hg19

STEP 1:
Load http://genome.ucsc.edu/



About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

06 July 2012 - hgText CGI To Be Retired

We are announcing the imminent retirement of the hgText CGI. hgText was the original version of the Table Browser, and was replaced by the new CGI, hgTables, in 2004. A very small number of our users are still accessing the database using the old hgText CGI. Beginning July 24, 2012, we will be removing the hgText CGI from our site. All users will need to use hgTables after that time.

19 June 2012 - Many New Assemblies Available in the Browser!

With the abundance of new vertebrate assemblies available in GenBank, the UCSC Genome Browser team has streamlined its browser release pipeline in the effort to keep pace. In the past week we have released a new browser on the Western painted turtle and updated assemblies for the cow, rat, and chicken. Browsers for the lesser hedgehog tenrec and the latest dog assembly should be available later this week, for a total of 18 assemblies released in the past half year. You can expect many new assemblies to show up on our website in upcoming months, along with broader multiple alignment annotations on selected organisms that include many of the new assemblies.

Keep in mind that some of these new assemblies may not yet be fully populated with annotation tracks. Some data, such as the data underlyi nmediately available, and other annotations (e.g., comparative genomics tracks) may take a while to generate. As always, our mandated focus is on vertebrates, with a particular emphas ne, although we do provide occasional updated assemblies for other selected model organisms (fly, worm, yeast).

We regret that we don't have the resources to support the many other genomics research communities who would like us to visualize their d u with information for setting up your own

Blue Vertical Tool Bar

Horizontal Tool Bar:

Genomes – genome browser

Blat – sequence search tool

Tables – table browser

Gene Sorter – gene based browser

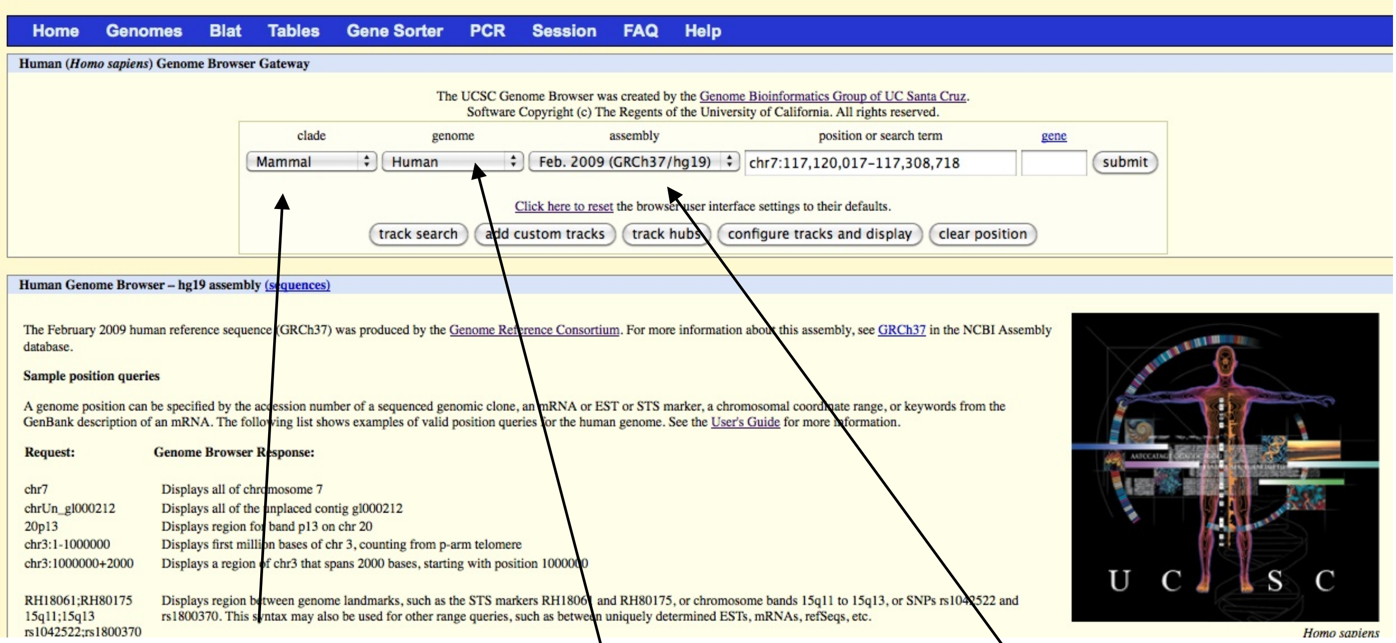
PCR – paired sequence search

VisiGene – in situ images browser

Proteome – protein browser

Session – manage session information

Help – user’s guide



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr7:117,120,017-117,308,718

Click here to reset the browser user interface settings to their defaults.

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.

Clade

Vertebrate
Deuterotsome
Insect
Nematode
Other

Genome

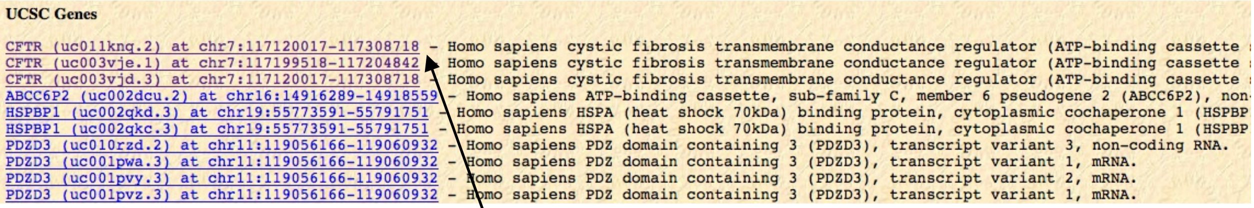
Human
Chimp
Orangutan
Rhesus
Marmoset, etc

Assembly

Dec. 2013 (GRCh38/hg38)
Feb. 2009 (GRCh37/hg19)
Mar. 2006 (NCBI36/hg18)
May2004(NCBI35/hg17)

Position: Can enter chromosome coordinates, cytogenetic band, gene name, STS marker, clone, text, range separated by “;”, i.e. RH18061;RH80175

STEP 3:
Enter “CFTR” in position / text window firstly and then gene window to see the different results.

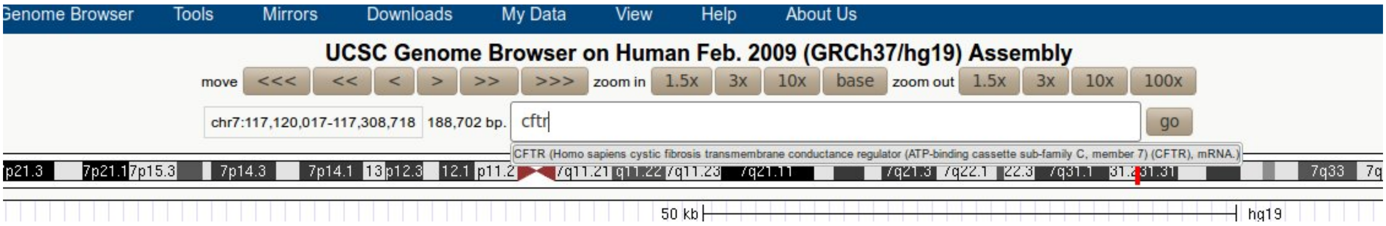


UCSC Genes

```
CFTR (uc011knq.2) at chr7:117120017-117308718 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette
CFTR (uc003vje.1) at chr7:117199518-117204842 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette
CFTR (uc003vjd.3) at chr7:117120017-117308718 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette
ABCC6P2 (uc002dcu.2) at chr16:14916289-14918559 - Homo sapiens ATP-binding cassette, sub-family C, member 6 pseudogene 2 (ABCC6P2), non
HSPBP1 (uc002qkd.3) at chr19:55773591-55791751 - Homo sapiens HSPA (heat shock 70kDa) binding protein, cytoplasmic cochaperone 1 (HSPBP
HSPBP1 (uc002qkc.3) at chr19:55773591-55791751 - Homo sapiens HSPA (heat shock 70kDa) binding protein, cytoplasmic cochaperone 1 (HSPBP
PDZD3 (uc010rzd.2) at chr11:119056166-119060932 - Homo sapiens PDZ domain containing 3 (PDZD3), transcript variant 3, non-coding RNA.
PDZD3 (uc001pwa.3) at chr11:119056166-119060932 - Homo sapiens PDZ domain containing 3 (PDZD3), transcript variant 1, mRNA.
PDZD3 (uc001pvy.3) at chr11:119056166-119060932 - Homo sapiens PDZ domain containing 3 (PDZD3), transcript variant 2, mRNA.
PDZD3 (uc001pvz.3) at chr11:119056166-119060932 - Homo sapiens PDZ domain containing 3 (PDZD3), transcript variant 1, mRNA.
```

STEP 4: Position result. This is a text search.
Click on the top link as this is the longest transcript.
“CFTR (uc003vjd.1) at chr7:117120017-117308718”
under UCSC Genes

STEP 5: Gene result
Takes you straight into the browser for the CFTR gene.
Obviously, this can only be used for known genes.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr7:117,120,017-117,308,718 188,702 bp. cftr

CFTR (Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) (CFTR), mRNA.)

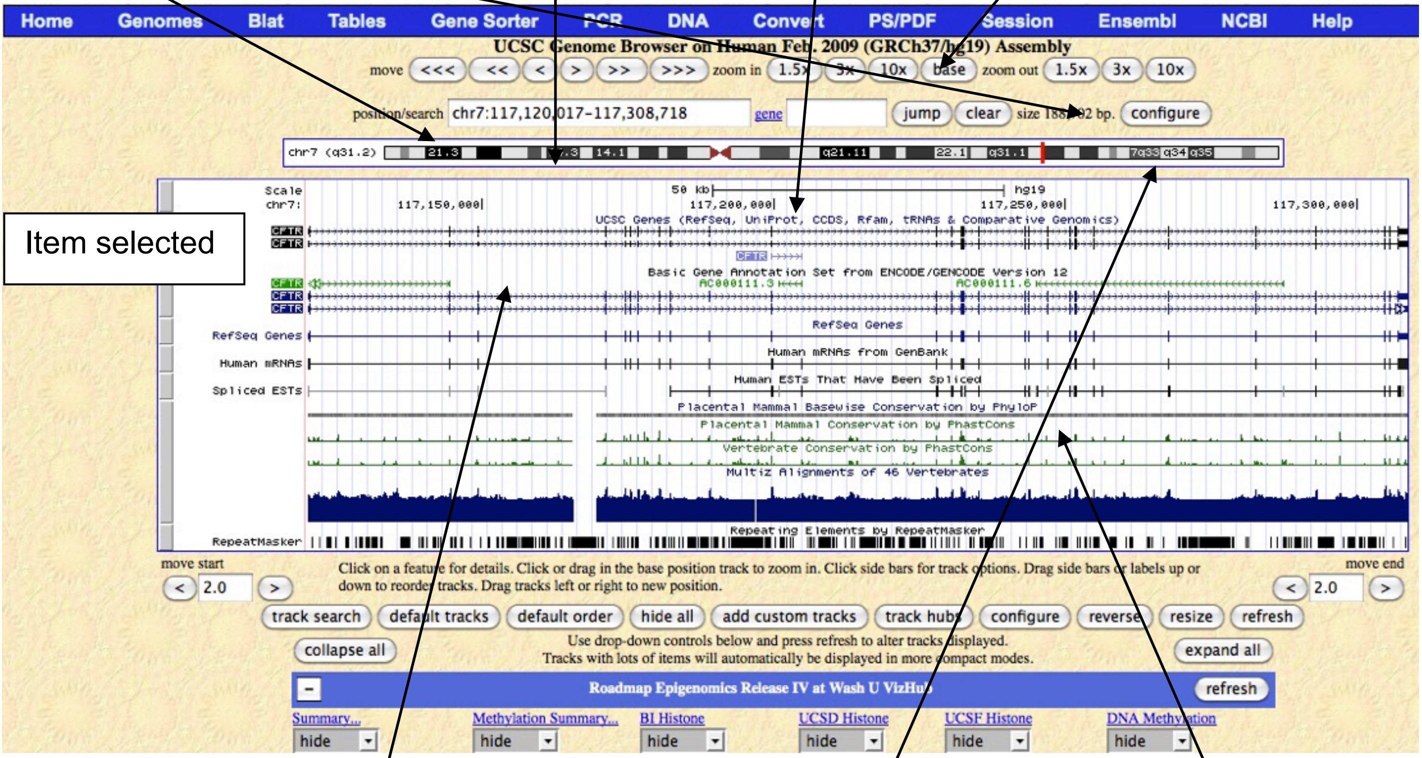
Navigation

Position text

Chromosome

Save graphic

Item selected


Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
PS/PDF
Session
Ensembl
NCBI
Help
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 chr7:117,120,017-117,308,718
gene
jump
clear
size 188,702 bp.
configure
chr7 (q31.2)
Scale
chr7:
50 kb
hg19
117,150,000
117,200,000
117,250,000
117,300,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CFTR
Basic Gene Annotation Set from ENCODE/GENCODE Version 12
AC000111.3
AC000111.6
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Spliced ESTs
Placental Mammal Basewise Conservation by PhyloP
Placental Mammal Conservation by PhastCons
Vertebrate Conservation by PhastCons
Multiz Alignments of 46 Vertebrates
Repeating Elements by RepeatMasker
RepeatMasker
move start
move end
2.0
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
collapse all
expand all
Roadmap Epigenomics Release IV at Wash U VizHub
Summary...
Methylation Summary...
BI Histone
UCSD Histone
UCSF Histone
DNA Methylation
hide

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Gene structure

Boxes – exons

(fat – translated, thin – UTR)

Lines – introns

Arrows indicate direction of transcription

Conservation "wiggle"

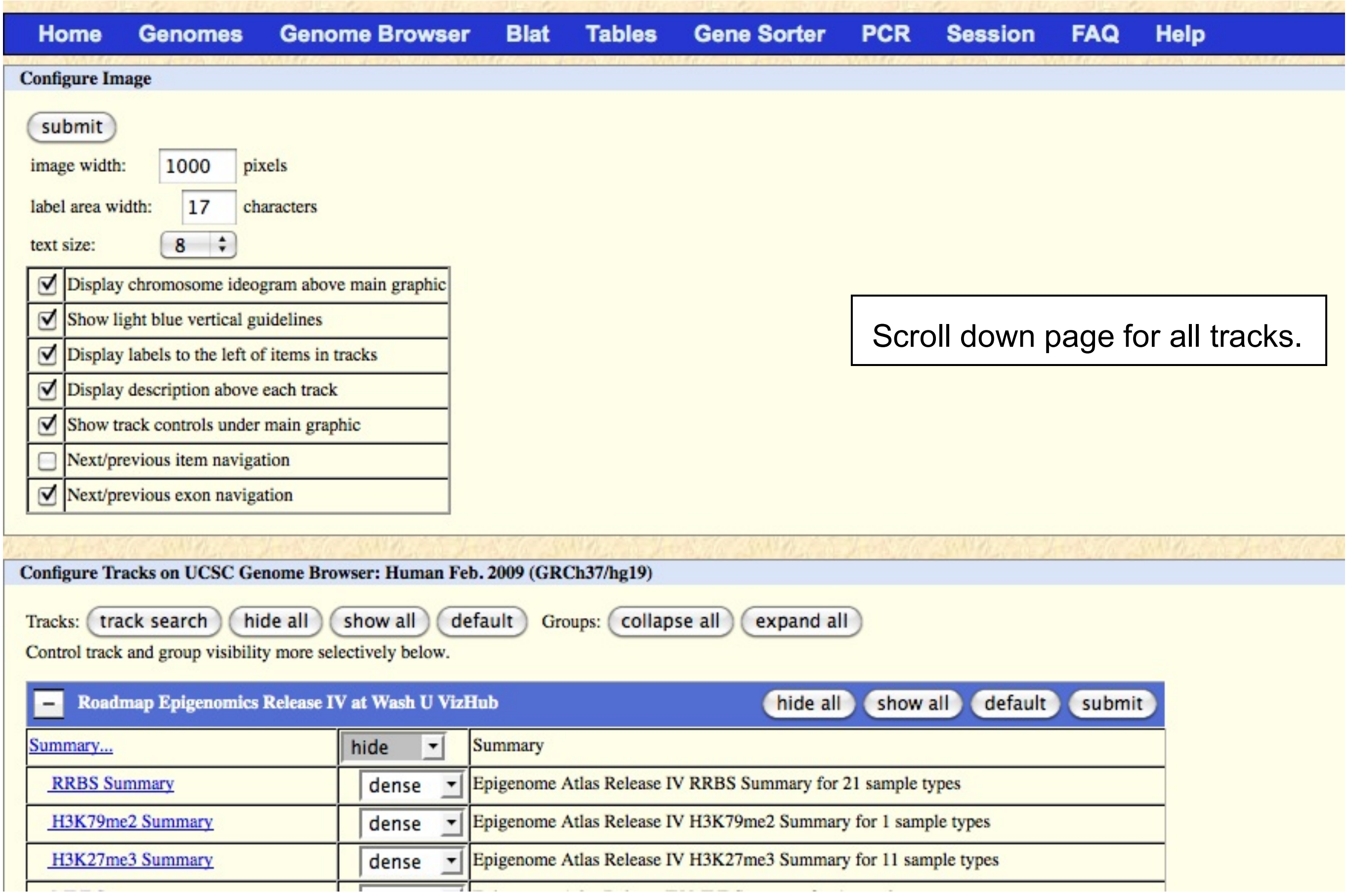
Annotations = "tracks"

Modes of view

- Hide (not show)
- Dense (all collapsed into 1 line)
- Squish (each item separate line but 50% height and packed)
- Pack (each item separate but efficiently stacked)
- Full (each item on separate line)

"DNA" link in horizontal tool bar allows you to retrieve the DNA sequence in this region

STEP 6:
Click on "Configure" button



Configure Image

submit

image width: 1000 pixels

label area width: 17 characters

text size: 8

☑ Display chromosome ideogram above main graphic

☑ Show light blue vertical guidelines

☑ Display labels to the left of items in tracks

☑ Display description above each track

☑ Show track controls under main graphic

☐ Next/previous item navigation

☑ Next/previous exon navigation

Scroll down page for all tracks.

Configure Tracks on UCSC Genome Browser: Human Feb. 2009 (GRCh37/hg19)

Tracks: track search | hide all | show all | default | Groups: collapse all | expand all

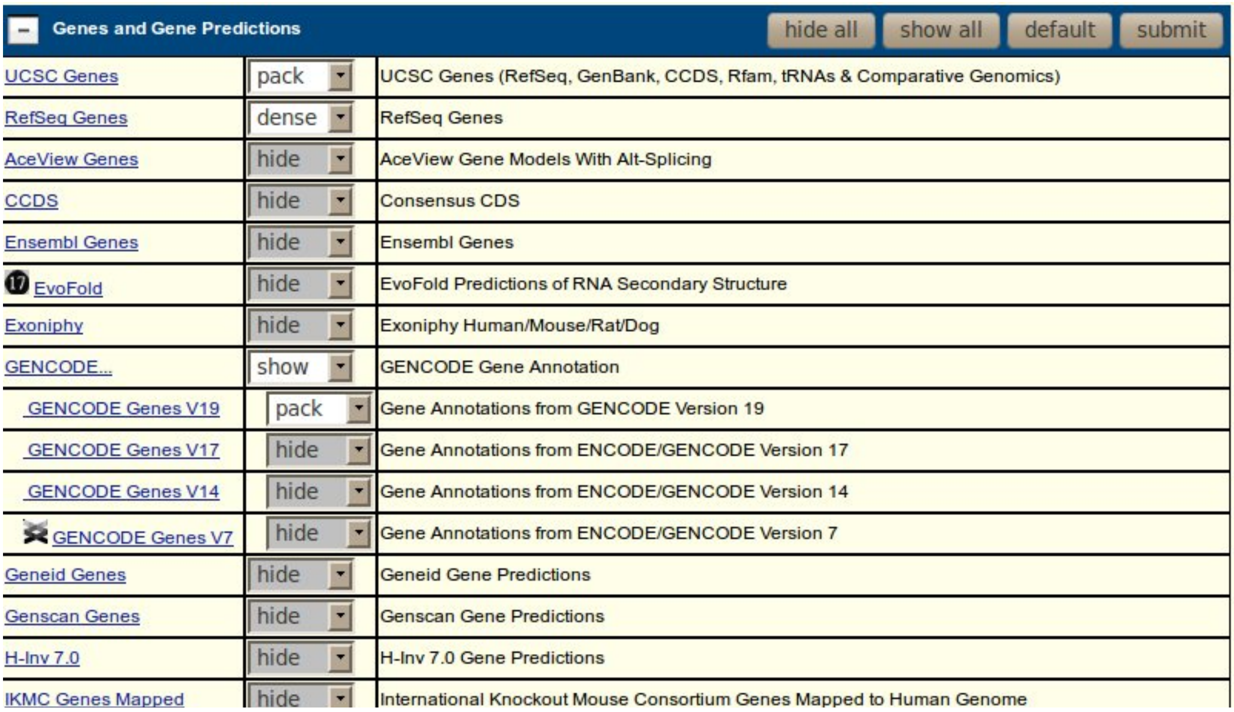
Control track and group visibility more selectively below.

− Roadmap Epigenomics Release IV at Wash U VizHub — hide all | show all | default | submit

Summary...	hide	Summary
RRBS Summary	dense	Epigenome Atlas Release IV RRBS Summary for 21 sample types
H3K79me2 Summary	dense	Epigenome Atlas Release IV H3K79me2 Summary for 1 sample types
H3K27me3 Summary	dense	Epigenome Atlas Release IV H3K27me3 Summary for 11 sample types

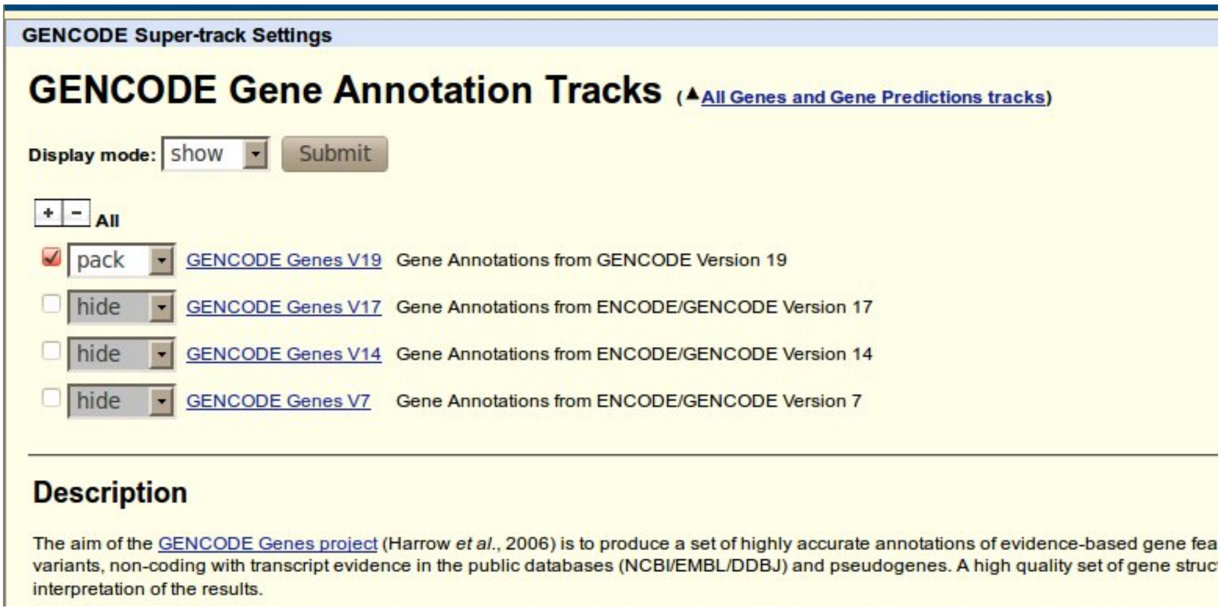
- Configure image size, text size, image labels
- Choose annotation “tracks” to display and display mode
- Tracks split into groups:
  - Mapping and Sequencing
  - Phenotype and Disease Associations
  - Genes and Gene Prediction
  - Literature
  - mRNA and EST
  - Expression
  - Regulation
  - Comparative Genomics
  - Variation and Repeats etc

STEP 7:
Scroll down to Gene and Gene Prediction Tracks.
Click on GENCODE.



Genes and Gene Predictions — hide all | show all | default | submit

Track	Display	Description
UCSC Genes	pack	UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes	dense	RefSeq Genes
AceView Genes	hide	AceView Gene Models With Alt-Splicing
CCDS	hide	Consensus CDS
Ensembl Genes	hide	Ensembl Genes
17 EvoFold	hide	EvoFold Predictions of RNA Secondary Structure
Exoniphy	hide	Exoniphy Human/Mouse/Rat/Dog
GENCODE...	show	GENCODE Gene Annotation
GENCODE Genes V19	pack	Gene Annotations from GENCODE Version 19
GENCODE Genes V17	hide	Gene Annotations from ENCODE/GENCODE Version 17
GENCODE Genes V14	hide	Gene Annotations from ENCODE/GENCODE Version 14
GENCODE Genes V7	hide	Gene Annotations from ENCODE/GENCODE Version 7
Geneid Genes	hide	Geneid Gene Predictions
Genscan Genes	hide	Genscan Gene Predictions
H-Inv 7.0	hide	H-Inv 7.0 Gene Predictions
IKMC Genes Mapped	hide	International Knockout Mouse Consortium Genes Mapped to Human Genome



GENCODE Super-track Settings

GENCODE Gene Annotation Tracks (▲All Genes and Gene Predictions tracks)

Display mode: show | Submit

+ − **All**

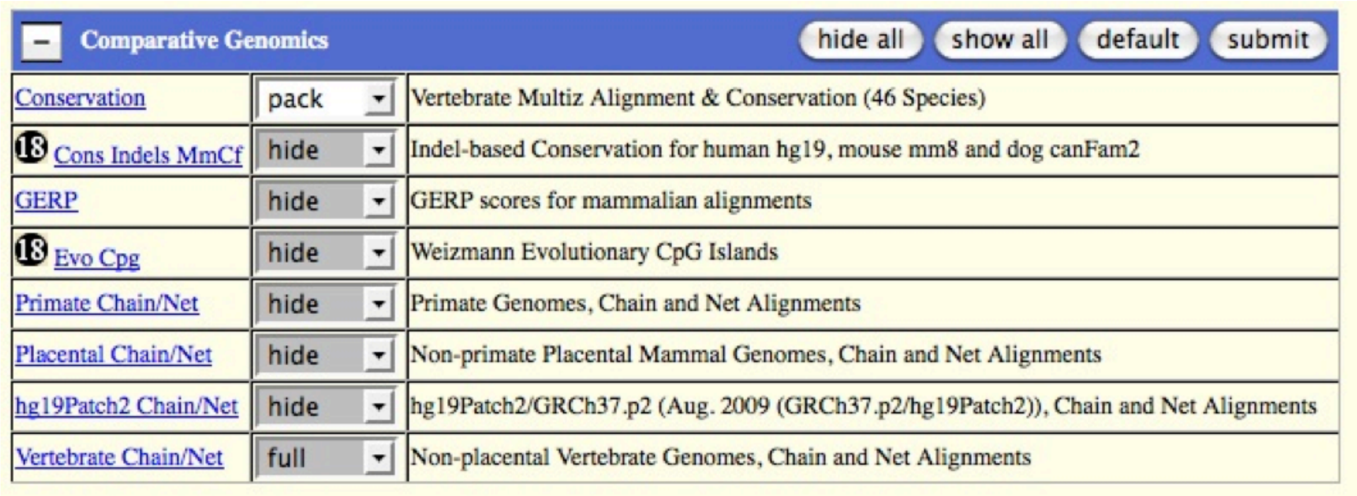
	Display	Track	Description
☑	pack	GENCODE Genes V19	Gene Annotations from GENCODE Version 19
☐	hide	GENCODE Genes V17	Gene Annotations from ENCODE/GENCODE Version 17
☐	hide	GENCODE Genes V14	Gene Annotations from ENCODE/GENCODE Version 14
☐	hide	GENCODE Genes V7	Gene Annotations from ENCODE/GENCODE Version 7

Description

The aim of the GENCODE Genes project (Harrow *et al.*, 2006) is to produce a set of highly accurate annotations of evidence-based gene fea variants, non-coding with transcript evidence in the public databases (NCBI/EMBL/DDBJ) and pseudogenes. A high quality set of gene struc interpretation of the results.

- Description of annotation including how it was created and who created it
- Many tracks have filters to restrict data displayed, or to color certain data a different color– this one allows you to include, exclude, or color markers from certain types of maps (individual genetic and RH STS maps)

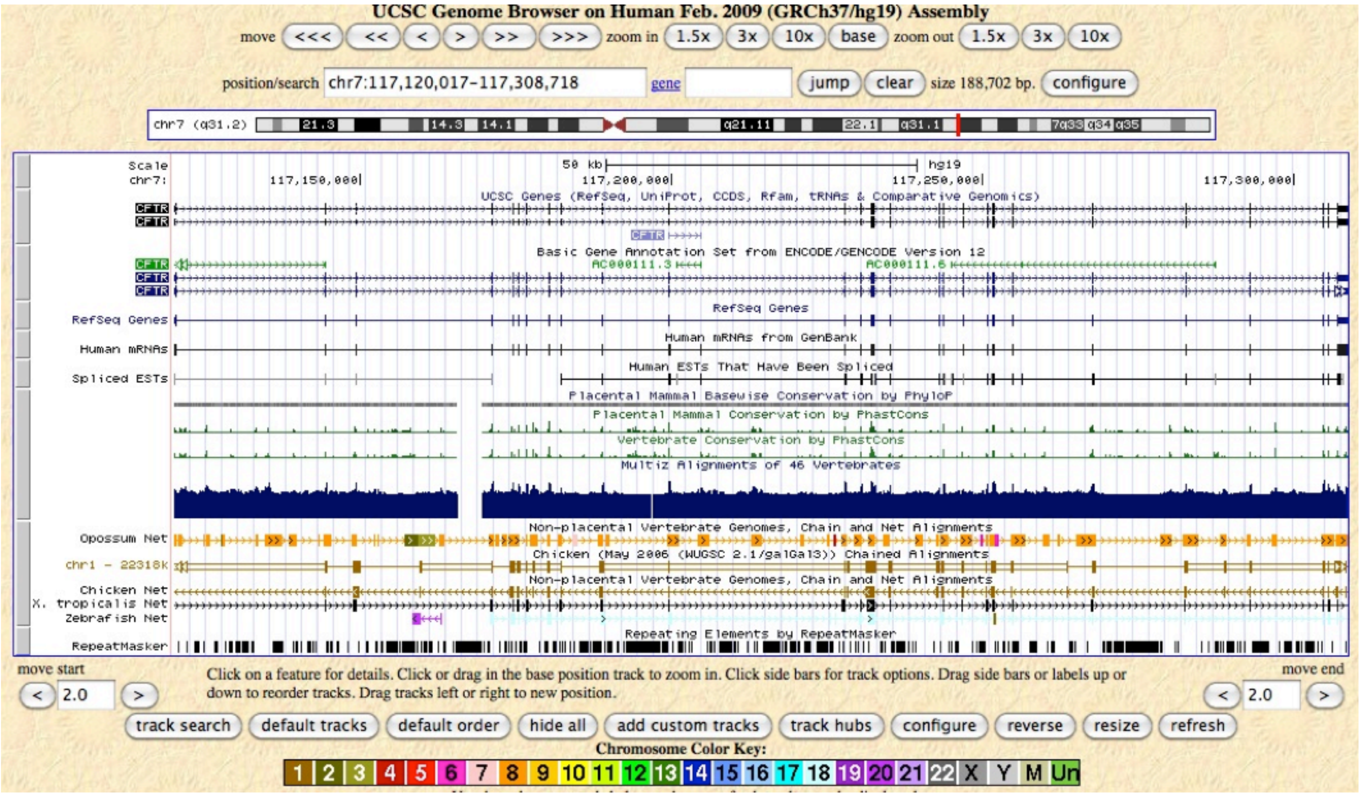
STEP 8:
Click on web browser “back” button to return to previous Configure screen.



Conservation	pack	Vertebrate Multiz Alignment & Conservation (46 Species)
18 Cons Indels MmCf	hide	Indel-based Conservation for human hg19, mouse mm8 and dog canFam2
GERP	hide	GERP scores for mammalian alignments
18 Evo Cpg	hide	Weizmann Evolutionary CpG Islands
Primate Chain/Net	hide	Primate Genomes, Chain and Net Alignments
Placental Chain/Net	hide	Non-primate Placental Mammal Genomes, Chain and Net Alignments
hg19Patch2 Chain/Net	hide	hg19Patch2/GRCh37.p2 (Aug. 2009 (GRCh37.p2/hg19Patch2)), Chain and Net Alignments
Vertebrate Chain/Net	full	Non-placental Vertebrate Genomes, Chain and Net Alignments

Boxes to left of track as well as links above track controls (below, bit shown) also bring up track description/filter page

STEP 9:
Scroll down to “Comparative Genomics”, Conservation, then click on “+” at the top of species selection.. Set Vertebrate Chain/Net track to “full”. Click on Submit button


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

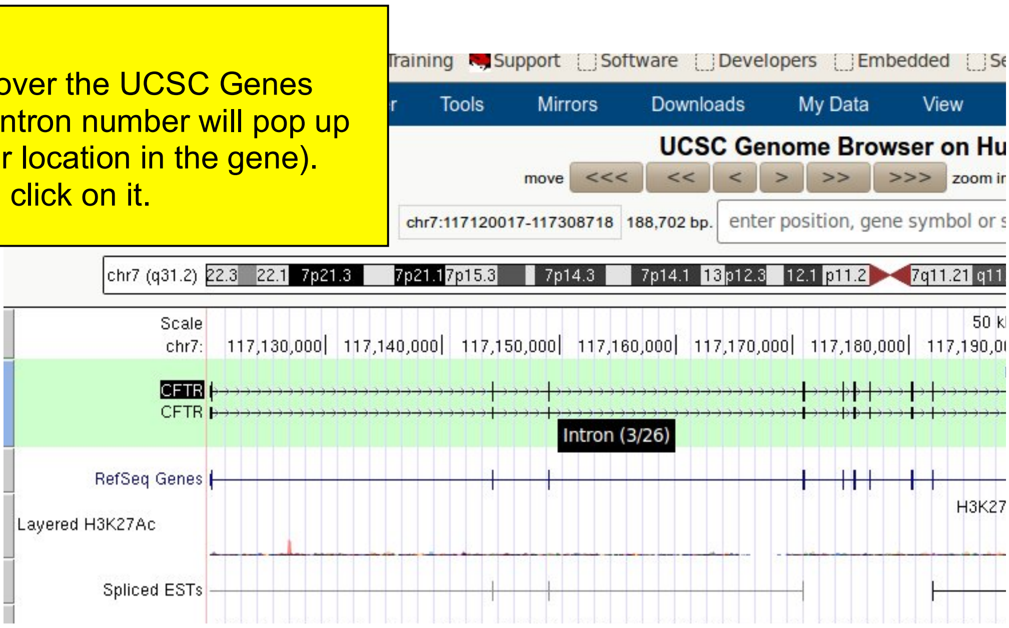
position/search chr7:117,120,017-117,308,718 size 188,702 bp.

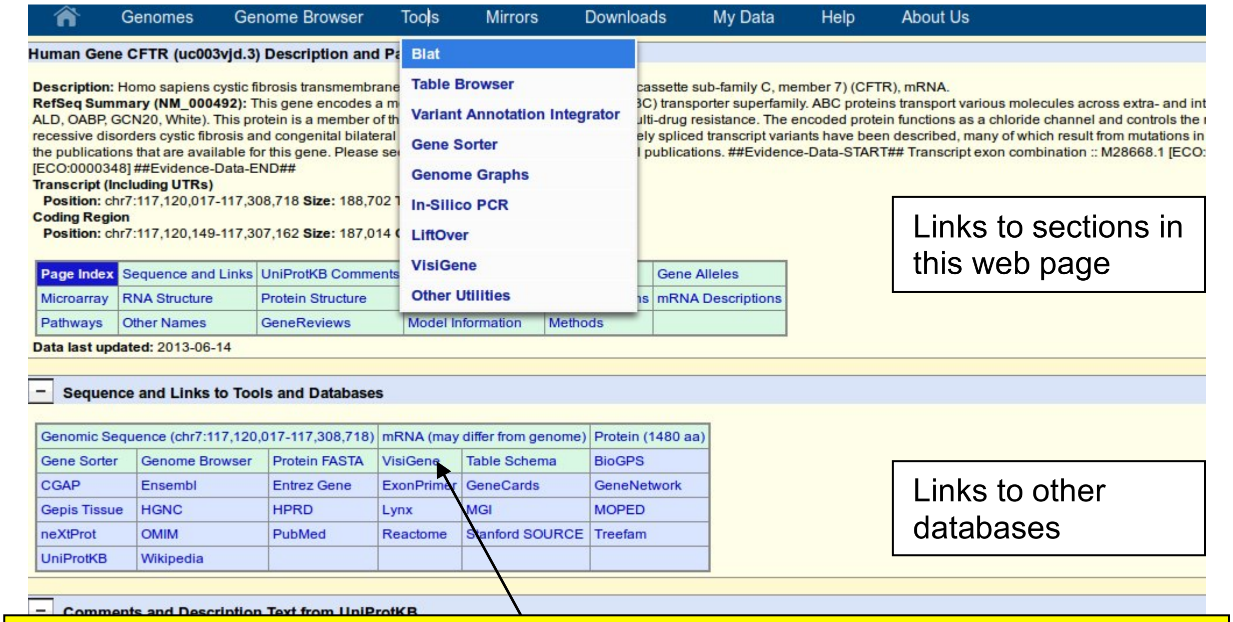
move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
gene
jump
clear
configure
chr7 (q31.2)
21.3
14.3
14.1
q21.11
22.1
q31.1
7q33 q34 q35
Scale
chr7:
50 kb
hg19
117,150,000
117,200,000
117,250,000
117,300,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CFTR
CFTR
CFTR
Basic Gene Annotation Set from ENCODE/GENCODE Version 12
AC000111.3
AC000111.6
CFTR
CFTR
CFTR
RefSeq Genes
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Placental Mammal Basewise Conservation by PhyloP
Placental Mammal Conservation by PhastCons
Vertebrate Conservation by PhastCons
Multiz Alignments of 46 Vertebrates
Opossum Net
Non-placental Vertebrate Genomes, Chain and Net Alignments
chr1 - 22318k
Chicken (May 2006 (WUGSC 2.1/galGal3)) Chained Alignments
Non-placental Vertebrate Genomes, Chain and Net Alignments
Chicken Net
X. tropicalis Net
Zebrafish Net
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
move end
< 2.0 >
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
Chromosome Color Key:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M Un

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Cross-species alignments are color-coded by chromosome. “Chain” shows all alignments, can overlap and “net” shows best 1-to-1 syteny mappings

STEP 10:
When you mouse over the UCSC Genes track, the exon or intron number will pop up (depending on your location in the gene). When this pops up click on it.





Human Gene CFTR (uc003vjd.3) Description and Pa

Description: Homo sapiens cystic fibrosis transmembrane … cassette sub-family C, member 7) (CFTR), mRNA.
RefSeq Summary (NM_000492): This gene encodes a m … 3C) transporter superfamily. ABC proteins transport various molecules across extra- and int
ALD, OABP, GCN20, White). This protein is a member of th … ulti-drug resistance. The encoded protein functions as a chloride channel and controls the
recessive disorders cystic fibrosis and congenital bilateral … ely spliced transcript variants have been described, many of which result from mutations in
the publications that are available for this gene. Please se … l publications. ##Evidence-Data-START## Transcript exon combination :: M28668.1 [ECO:
[ECO:0000348] ##Evidence-Data-END##
Transcript (Including UTRs)
Position: chr7:117,120,017-117,308,718 **Size:** 188,702 T
Coding Region
Position: chr7:117,120,149-117,307,162 **Size:** 187,014 C

Page Index	Sequence and Links	UniProtKB Comments			Gene Alleles
Microarray	RNA Structure	Protein Structure		s	mRNA Descriptions
Pathways	Other Names	GeneReviews	Model Information	Methods	

Data last updated: 2013-06-14

Links to sections in this web page

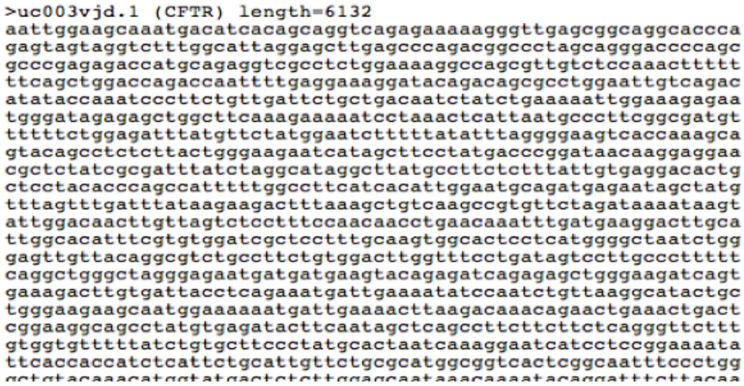
Sequence and Links to Tools and Databases

Genomic Sequence (chr7:117,120,017-117,308,718)			mRNA (may differ from genome)		Protein (1480 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	HGNC	HPRD	Lynx	MGI	MOPED
neXtProt	OMIM	PubMed	Reactome	Stanford SOURCE	Treefam
UniProtKB	Wikipedia				

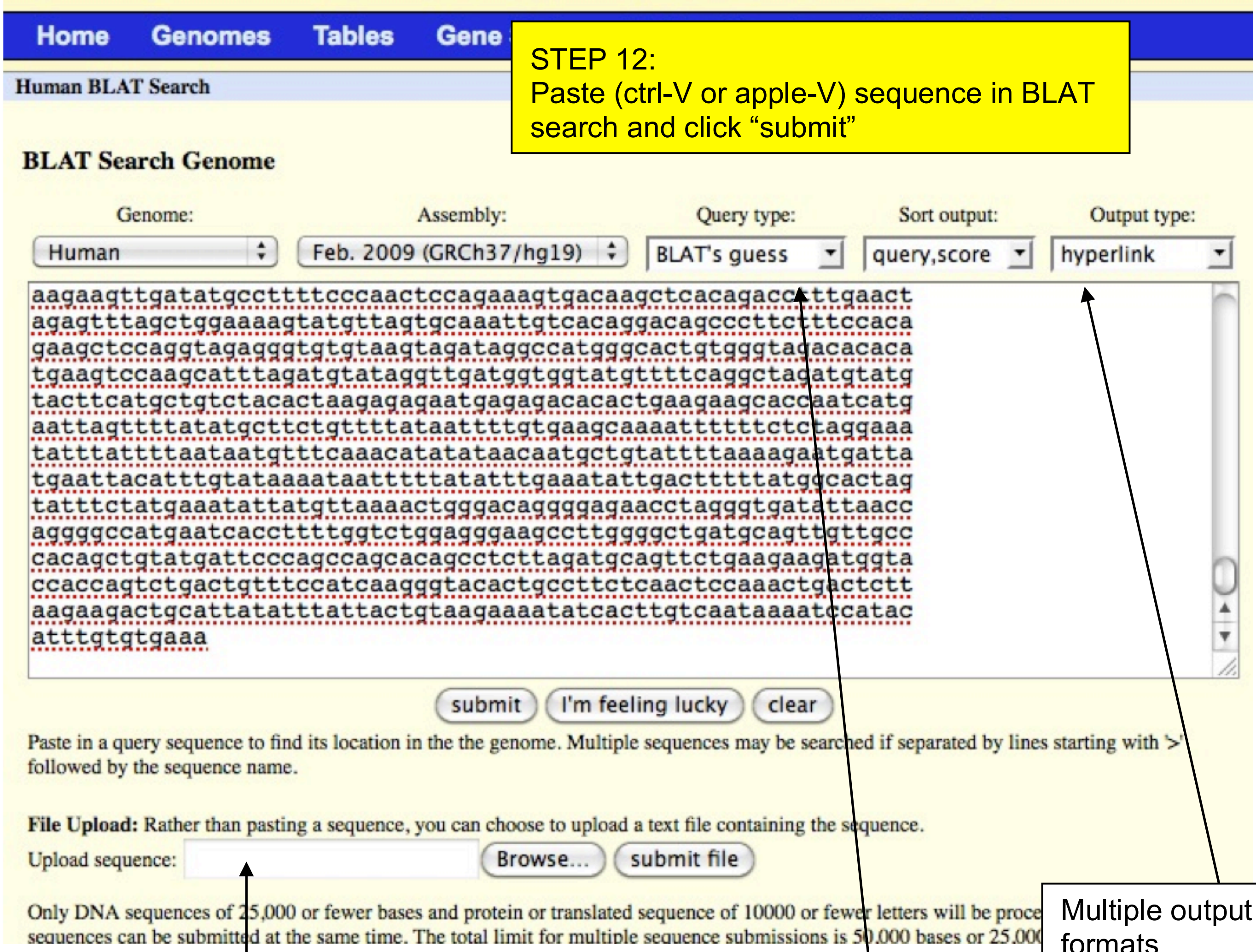
Links to other databases

Comments and Description Text from UniProtKB

STEP 11:
Click on “mRNA” sequence link. In the new window that displays FASTA sequence, select all and copy sequence. Return to above screen (back button in web browser) and click on Tools on the horizontal blue bar at top, the select BLAT.



```
>uc003vjd.1 (CFTR) length=6132
aattggaagcaaatgacatcacagcaggtcagagaaaaagggttgagcggcaggcaccca
gagtagtaggtctttggcattaggagcttgagcccagacggccctagcagggaccccagc
gcccgagagaccatgcagaggtcgcctctggaaaaggccagcgttgtctccaaacttttt
ttcagctggaccagaccaattttgaggaaaggatacagacagcgcctggaattgtcagac
atataccaaatcccttctgttgattctgctgacaatctatctgaaaaattggaaagagaa
tgggatagagagctggcttcaaagaaaaatcctaaactcattaatgcccttcggcgatgt
tttttctggagatttatgttctatggaatctttttatatttaggggaagtcaccaaagca
gtacagcctctcttactgggaagaatcatagcttcctatgacccggataacaaggaggaa
cgctctatcgcgatttatctaggcataggcttatgccttctctttattgtgaggacactg
ctcctacacccagccatttttggccttcatcacattggaatgcagatgagaatagctatg
tttagtttgatttataagaagactttaaagctgtcaagccgtgttctagataaaataagt
attggacaacttgttagtctcctttccaacaacctgaacaaatttgatgaaggacttgca
ttggcacatttcgtgtggatcgctcctttgcaagtggcactcctcatggggctaatctgg
gagttgttacaggcgtctgccttctgtggacttggtttcctgatagtccttgcccttttt
caggctgggctagggagaatgatgatgaagtacagagatcagagagctgggaagatcagt
gaaagacttgtgattacctcagaaatgattgaaaatatccaatctgttaaggcatactgc
tgggaagaagcaatggaaaaaatgattgaaaacttaagacaaacagaactgaaactgact
cggaaggcagcctatgtgagatacttcaatagctcagccttcttcttctcagggttcttt
gtggtgtttttatctgtgcttccctatgcactaatcaaaggaatcatcctccggaaaata
ttcaccaccatctcattctgcattgttctgcgcatggcggtcactcggcaatttccctgg
gctgtacaaacatggtatgactctcttggagcaataaacaaaatacaggatttcttacaa
```



Human BLAT Search

BLAT Search Genome

Genome: Human
Assembly: Feb. 2009 (GRCh37/hg19)
Query type: BLAT's guess
Sort output: query,score
Output type: hyperlink

```
aagaagttgatatgccttttcccaactccagaaagtgacaagctcacagaccattgaact
agagtttagctggaaaagtatgttagtgcaaattgtcacaggacagcccttctttccaca
gaagctccaggtagagggtgtgtaagtagataggccatgggcactgtgggtagacacaca
tgaagtccaagcatttagatgtataggttgatggtggtatgttttcaggctagatgtatg
tacttcatgctgtctacactaagagagaatgagagacacactgaagaagcaccaatcatg
aattagttttatatgcttctgttttataattttgtgaagcaaaatttttctctaggaaa
tatttattttaataatgtttcaaacatatataacaatgctgtattttaaaagaatgatta
tgaattacatttgtataaaataattttatatttgaaatattgacttttatggcactag
tatttctatgaaatattatgttaaaactgggacagggagaacctagggtgatattaacc
aggggccatgaatcacctttggtctggagggaagccttggggctgatgcagttgttgcc
cacagctgtatgattcccagccagcacagcctcttagatgcagttctgaagaagatggta
ccaccagtctgactgtttccatcaagggtacactgccttctcaactccaaactgactctt
aagaagactgcattatatttattactgtaagaaaatatcacttgtcaataaaatccatac
atttgtgtgaaa
```

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence:

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be proce sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000

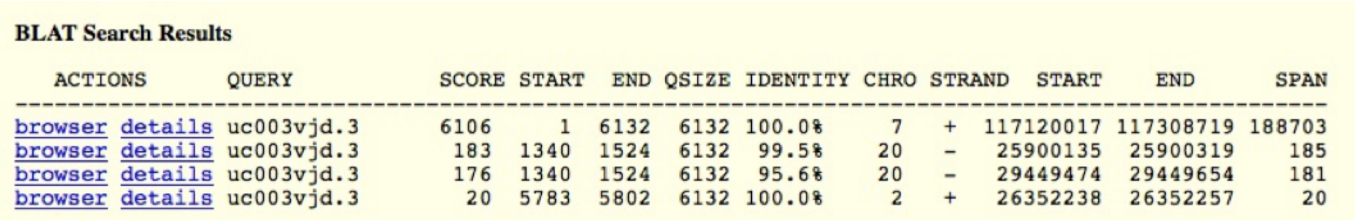
STEP 12:
Paste (ctrl-V or apple-V) sequence in BLAT search and click "submit"

Multiple output formats

Can upload a file containing sequence

Aligns dna, protein, translated RNA/DNA sequences

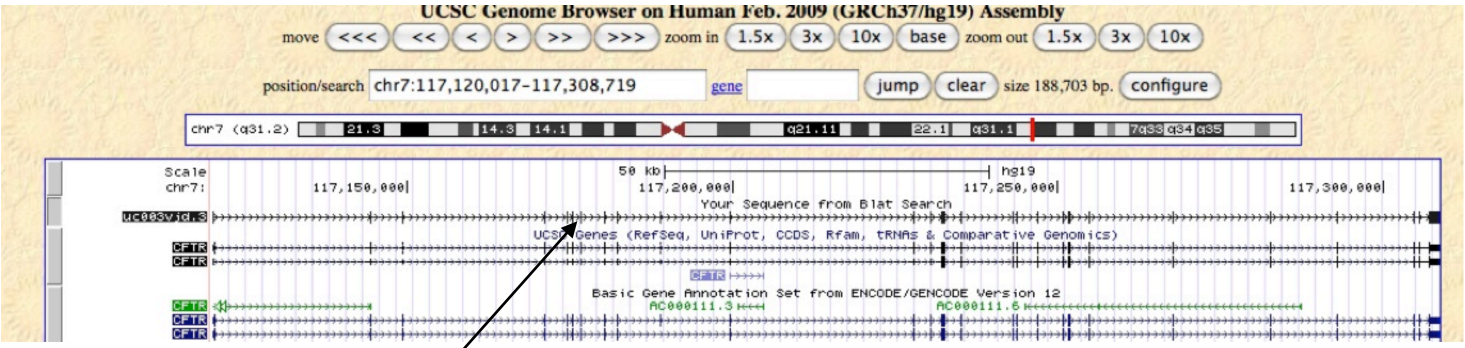
STEP 13:
Click on top "browser" link



BLAT Search Results

```
   ACTIONS        QUERY              SCORE START   END QSIZE IDENTITY CHRO STRAND   START      END      SPAN
-------------------------------------------------------------------------------------------------------------
browser details uc003vjd.3          6106     1  6132  6132 100.0%      7    +   117120017 117308719 188703
browser details uc003vjd.3           183  1340  1524  6132  99.5%     20    -    25900135  25900319    185
browser details uc003vjd.3           176  1340  1524  6132  95.6%     20    -    29449474  29449654    181
browser details uc003vjd.3            20  5783  5802  6132 100.0%      2    +    26352238  26352257     20
```

Details of the alignments



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 chr7:117,120,017-117,308,719 gene jump clear size 188,703 bp. configure
chr7 (q31.2) 21.3 14.3 14.1 q21.11 22.1 q31.1 7q33 q34 q35
Scale 50 kb hg19
chr7: 117,150,000 117,200,000 117,250,000 117,300,000
Your Sequence from Blat Search
uc003vjd.3
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CFTR
CFTR
CFTR
Basic Gene Annotation Set from ENCODE/GENCODE Version 12
CFTR
AC000111.3
AC000111.6
CFTR
CFTR

Aligned sequence from BLAT search displayed as a track.

Additional UCSC Exercises

1.Exploring features related to a gene

Find the human HRAS gene in the February 2009 assembly.

How many variants are displayed for this gene in the UCSC Genes track? Does the number differ in the GENODE genes track? What do the different colours of the transcripts mean?

What is its molecular function according to GO annotation? What signalling pathways is it associated with? With what diseases is HRAS associated?

Hint: click on the gene to get to the detail page, and search in this page for this information.

How many amino acids does the HRAS transcript code for? Which Pfam domains does the protein product contain?

In which chromosomal band, on which clone and contig in the genomic sequence assembly is HRAS located?

Hint: in the main browser screen, open tracks associated with clones (BAC End Pairs, Assembly) and contigs (Map Contigs) to "pack" or "full" display modes.

Look at regions that are conserved with mouse (turn on Placental Chain/Net in Comparative Genomics, and Other RefSeq in Genes and Gene Prediction tracks).

Is there a putative mouse (*Mus)* ortholog in the Other RefSeq track? From the chains and nets, can you find where is it in the mouse genome?

2. Exploring a region

Display the region between markers D12S764 and D12S1871 in the February 2009 human assembly in the main browser display.

Hint: use the marker names separated by a ";" in the position box.

How many clones/contigs are used to make this portion of the assembly?

Zoom in on the TENC1 gene by drawing a box around it with holding down the left mouse button.

Which STS markers are contained within the 3' UTR of the TENC1 gene? How many synonyms does each marker have? How many RefSeq mRNAs are displayed for this gene?

Hint: make sure the tracks are in "pack" or "full" display mode.

Zoom in on the TENC1 gene. Identify the coding SNPs. How many are synonymous, frameshift and missense?

Hint: turn on the Common SNP track. Go to the filtering options of the "Function" category. De-select all but "Synonymous variant", "frameshift variant" and "missense variant". Notice that coding SNPs are coloured red (Non-Synonymous i.e missense and frameshift variants) and green (Synonymous variants).

3. Exploring the mouse genome

Hint: use what you know about navigation in the human browser

Go to the main "Genomes" page and select the Dec 2011 mouse assembly. Bring up a display of mouse chromosome 2 between 18,500,000 and 18,900,000.

How many UCSC genes are predicted in this region? For one of the UCSC genes, find some information about its function, and look at an entry for it in Entrez Gene, UniProt or the Jackson Lab's Mouse Genome Informatics site.

Make sure the conservation track is open to "full". Zoom in on the gene Commd3. Which of the species in this track does not show any conservation with mouse in the region that is spanned by the first exon of this gene? Why may this be?

4. Other functions

In the Mouse genome, mouse over the Tools link on the horizontal tool bar to reveal the menu and select In-Silico PCR. Input GAATAGGGGAGTTAGAGGGGG as the Forward Primer, and GAAACTCTTTTTTTTTCTTCAGTGTG as the Reverse Primer, and search in the Mouse genome. What are the primer melting temperatures for the two primers? What is the name of the genomic clone that these primers correspond to, what type is it and how long is it? What Repeat Elements span this marker? Would this marker be considered a microsatellite?

Exercise Answers

1. The HRAS gene is located at chr11:532,242-535,550 in the February 2009 sequence. The UCSC Genes track displays four variants and the GENCODE v14 genes track has eleven variants. The UCSCS track had black and dark blue transcripts. Black transcripts have an entry in the Protein Data Bank (PDB) and dark blue transcripts have been reviewed. The GENCODE genes are dark blue and green. Dark blue are coding and green are non-coding genes.

Clicking the first black transcript in the UCSC gene track (uc001lpv.3), you can see that GO lists the Molecular Functions of HRAS as nucleotide binding, protein binding, GTP binding and protein C-terminus binding. The KEGG database states that HRAS participates in many different pathways including the MAPK and VEGF signalling pathways. BioCarta from the NCI Cancer Genome Anatomy Project includes HRAS in many pathways including EGF signalling pathway and T cell receptor signalling pathway. HRAS has been implicated in several diseases including Costello's syndrome and several cancers. OMIM is a good resource for additional disease information.

This same transcript encodes 189 amino acids (found in Sequence and links to Tools and Databases). HRAS contains the InterPro domains Small_GTP-bd_dom, Small_GTPase and Small_GTPase_Ras protein domain (found in Protein Structure Information).

HRAS is in band 11p15.5, in contig GL000101.1, and in clone AC137894.5 in the sequence assembly. The BAC End Pairs track additionally shows it is contained in clones RP11-1007G14, RP11-392J11, and RP11-1021K7.

The putative mouse ortholog according to the Other RefSeq track is Mus Hras1. Clicking on the Mouse Chain (pink bar at top of track) shows this region is part of a ~3Mb syntenic region on mouse chr7:140,845,391-143,785,351. Clicking on the Open Mouse Browser link on this detail page brings up the region in the Mouse browser. The RefSeq track shows the Hras1 gene. Hras1 itself is at chr7:141,189,934-141,194,004. This gene can also be found by opening the Mouse Browser and typing Hras1 into the gene or position box.

2. All or parts of 4 clones (AC107016.20, AC107202.14, AC068888.35, AC073573.27) were used to sequence the genome between markers D12S764 and D12S1871.

The STS markers RH52721 and RH44510 are contained in the 3'UTR of the TENC1 gene. Synonyms (Other Names) for RH52721 are WI-21011, HSA.21006, and STS-R05823 and for RH44510 is STSG4946 as seen on the detail pages for these markers. There are 3 RefSeq mRNAs for TENC1.

There are eight coding SNPs identified in TENC1 – rs12369033, rs2293062, rs11170389, rs11558984, rs34044566, rs73099915, rs118159776, and rs12816417. Five of them are synonymous SNPs (coloured green) and three are non-synonymous (coloured red). The colourings for types of SNPs can be changed on the description page for this track.

3. There are four UCSC Genes in the Mouse genome between 18,500,000 and 18,900,000. They are Commd3, Bmi1 (2 isoforms), BC061194 and Pip4k2a. Information about these genes will vary with the gene selected and the resource used.

The chicken genome does not show any conservation. There are several reasons why this might be. Most likely, either the chicken genome is missing the sequence for this region in its assembly (still needs to be sequenced), or this region was lost at some point in evolution by the chicken or one of its ancestors. The corresponding region in the chicken genome does show unsequenced gaps in this region as can be seen by opening the Chicken Net track (found under Vertebrate Chain/Net) and following links on the detail page, and this is most likely the cause. Several other genomes also have missing sequence, including sheep and armadillo which are shown as a pale yellow line, which indicates that the species has N's in the gap region. This reflects uncertainty in the relationship between the DNA of both species, due to lack of sequence in the relevant portions of the aligning species. A few other species have a double black line, such as tree-shrew and zebra finch, which indicates the one or more bases in the region do not align.

4. The initial results page from the PCR function displays melting temperatures for the primers:

Forward: 60.1 C gaataggggagttagagggg

Reverse: 59.0 C gaaactctttttttttcttcagtgtg

The temperature calculations are done assuming 50 mM salt and 50 nM annealing oligo concentration. The code to calculate the melting temp comes from Primer3.

These primers correspond a region on clone ID GL456090.2, it is type is F (finished) and it is 19375614bp in length. According to the Repeat Masker track, there are two LINE elements and a simple repeat in this region. The simple repeat is a TG/CA di-nucleotide repeat. This marker is considered a microsatellite, defined as tandemly repeated DNA usually shorter than 150 bases with repeat unit lengths less than about 10 bases. Switch on the microsatellite track to confirm this.