GENOME BROWSER (course booklet pages 42-44)

Answer 1 - Exploring a genomic region in human using the Ensembl Browser

a) Go to the Ensembl homepage (http://www.ensembl.org/) and search for '13:31873863-32623863' in the human species page in Ensembl.

This genomic region is located on cytogenetic band q13.1.
It is made up of seven contigs, indicated by the alternating light and dark blue coloured bars in the 'Contigs' track.

b) Draw with your mouse a box around the BRCA2 transcripts. And then click on 'Jump to region' in the pop-up menu.

c) Click 'Configure this page' in the side menu and type 'tilepath' in the 'Find a track' text box. Select 'Tilepath' and save and close by clicking on '✓'.

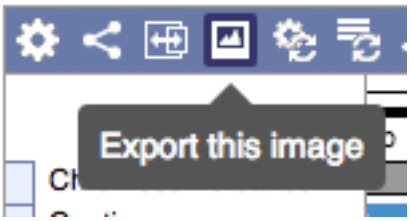
Two BAC clones, not only a single one, contain the complete *BRCA2* gene. Clone RP11-37E23 contains most of the gene, but not its very 3' end. Click on the second clone to find out its ID.

d) Click 'Configure this page' in the side menu, type 'refseq' in the 'Find a track' text box, select 'RefSeq human import' and choose the track style 'Expanded with labels'. Save and close.

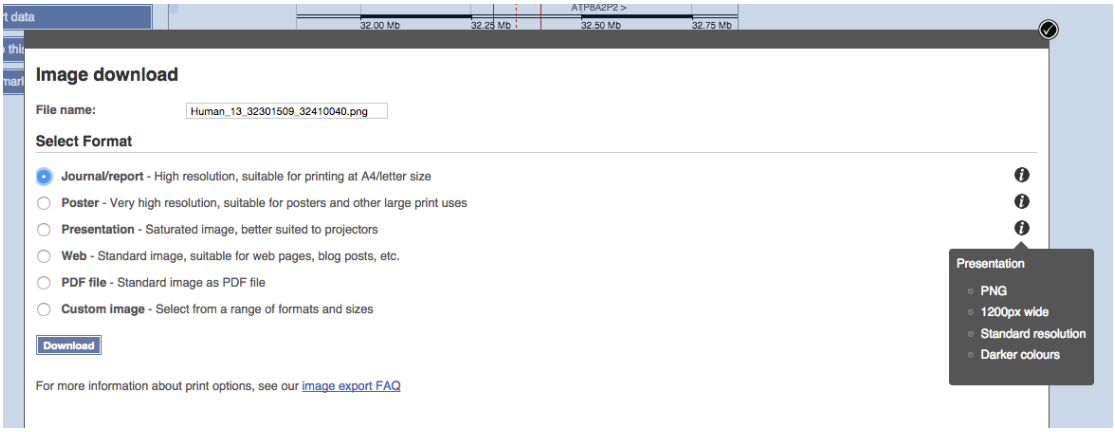
Click on individual transcript models to retrieve more information about them.

There have been four transcripts annotated by RefSeq for the *BRCA2* gene. Three have got an XM prefix and one has got the NM prefix. The former are predictions, whereas the latter is a curated transcript. Let's compare NM_000059.3 to the Ensembl transcript BRCA2-001 (ENST00000380152). Both encode a 3,418 aa protein. Look at the differences at the UTRs when comparing the Ensembl transcript to the RefSeq transcript.

e) Click on the 'Export this image' icon on the panel of the image you want to export



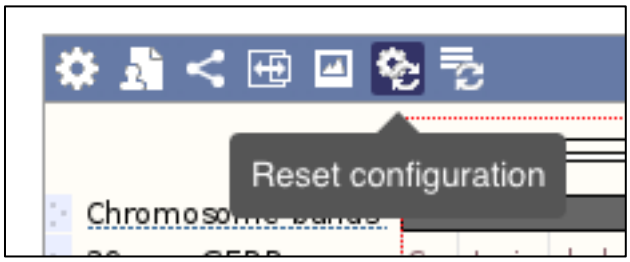
All options are PNG and they vary in their resolution. To download a picture in Ensembl to use in your publication please select the option 'Journal/report'.



f) Click 'Export data' in the side menu, then on 'Next>' and on 'Text'. There are different options for sequence export in FASTA e.g. unmasked, repeat masked (soft), etc.

Note that the exported sequence has a header line that provides information about the genome assembly (e.g. GRCh38), the chromosome, the start and end coordinates and the strand (1 for forward, -1 for reverse).

g) Click 'Configure this page' in the side menu and then on 'Reset configuration'. Save and close. Or click on the icon at the top of the image as illustrated below:



Answer 2 - Exploring a gene and its transcripts in human

a) The human *F9* gene is located on the X chromosome on the forward strand. Three transcripts were annotated for this gene, namely ENST00000218099 (F9-001), ENST00000394090 (F9-201) and ENST00000479617 (F9-002).

b) The longest transcript is ENST00000218099 (F9-001). The length of this transcript is 2780 base pairs and the length of the encoded protein is 461 amino acids.

Click on the transcript 'F9-001' in the 'Gene Summary' display.

It is an Ensembl/Havana merge transcript, annotated by both automatic and manual modes of annotation. Golden transcripts are from the Ensembl/Havana merged gene set. The gene set you see in Ensembl is the GENCODE set of genes, a set of high and comprehensive annotation used by several different projects such as 1000 Genomes.

Click on the Ensembl Transcript ID 'ENST00000218099' in the list of transcripts.

This transcript has got eight exons.

Click on 'Exons' in the side menu.

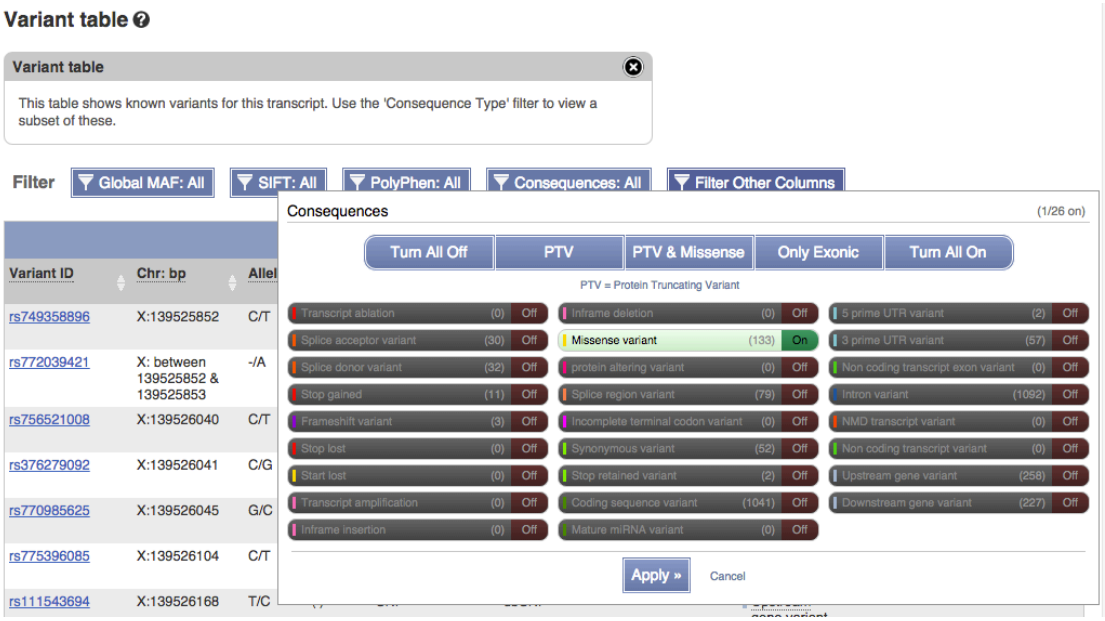
The first and last exons are partially untranslated (sequence shown in orange).

c) Now click on 'Protein summary' in the side menu.

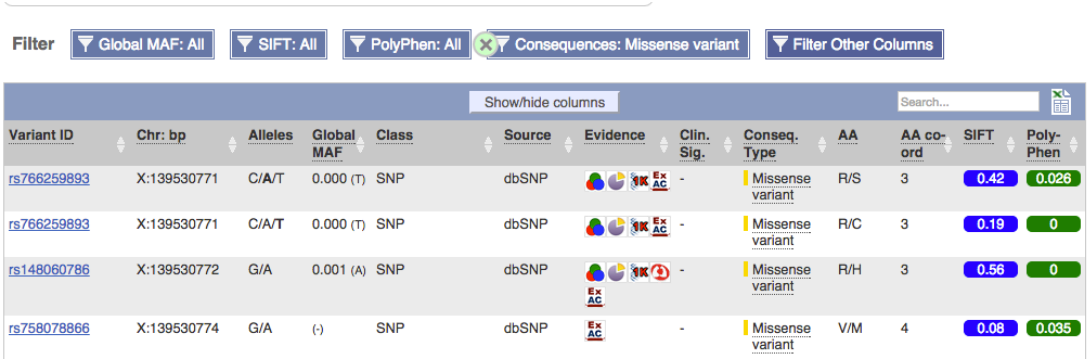
The peptidase activity of the protein resides in the peptidase domain that is located in the C-terminal portion of the protein (which

corresponds to the 3' end of the nucleotide sequence). The protein does not seem to contain any transmembrane domains according to current data available.

d) Click on 'Variant table' in the side menu. You may want to filter this table, as there are too many entries in there. Turn all the consequence types off, turn 'missense variants' on and click on 'Apply'.



You will see a table similar to the one below:

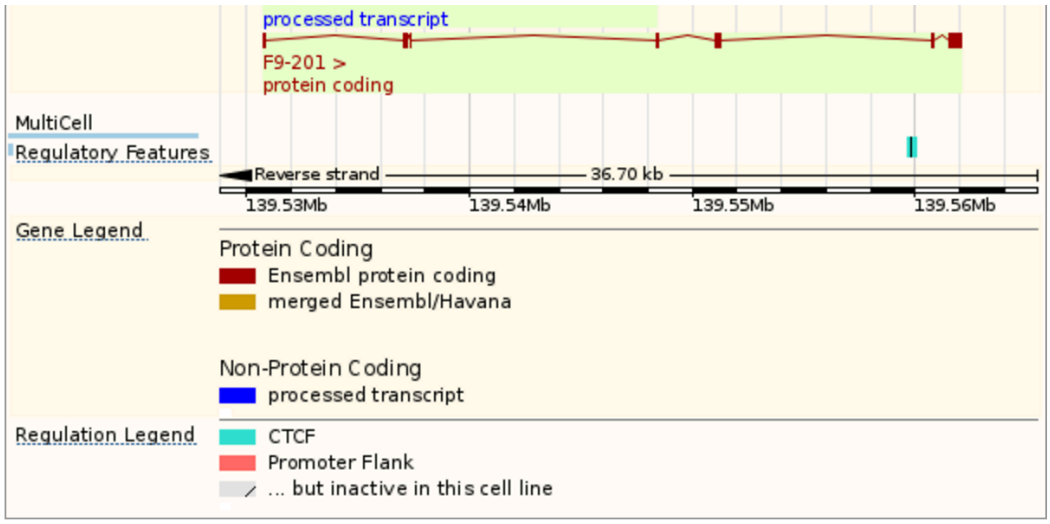


For the protein encoded by ENST00000218099 many missense variants have been mapped to. They are all imported from dbSNP (i.e. all variants with an identifier starting with 'rs').

Look at the code in the Evidence column of the table to find the data supporting the evidence.

Note that the number stated in the table corresponds to the number of variant consequences, not the number of variants.

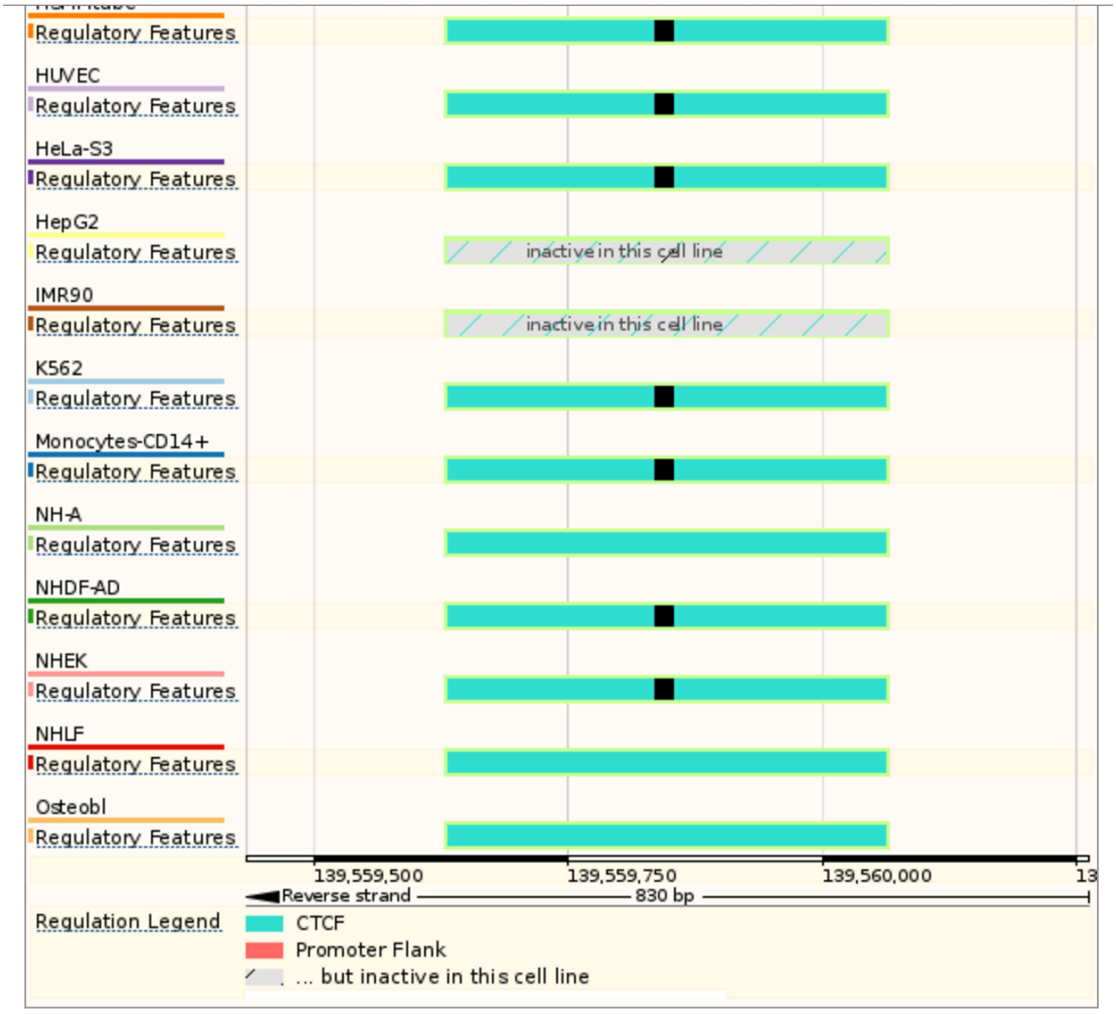
e) Click on 'Regulation' in the side menu. There is just one type of Regulatory feature annotated in the region of this gene.



Show/hide columns | Filter

Reg. region	Analysis	Type	Location	Length (bp)	Sequence (pc
ENSR00001491891	Ensembl Regulatory Build	CTCF Binding Site	X:139559632-139560061	472	TCTAGGACATA ACCATCTGATA CATGCCTAAGA GAAAAGGATTT CTATAAAGGTC AGACATTGAGA GAGATAATGGA CACTTTGCCC

Click on the ID ENSR00001491891 to go to the Regulation tab and find out this feature is active in most cells but HepG2 and IMR90:



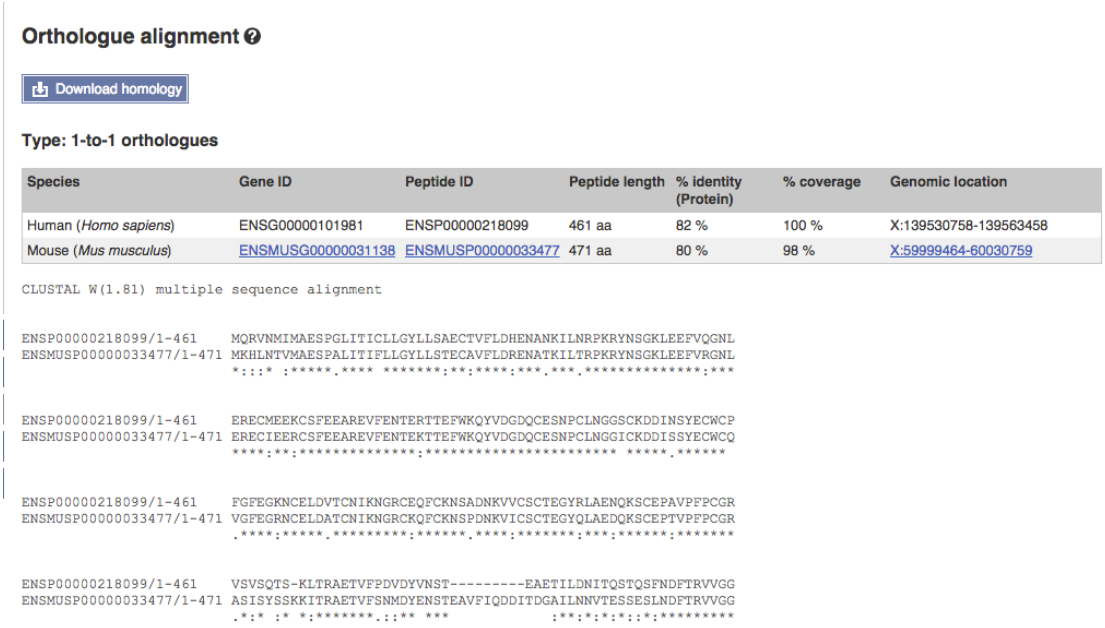
f) Click on 'Orthologues' in the side menu. We have identified genes in seven rodents that are orthologous to the human *F9* gene. No orthologues have been annotated in pika (*Ochotona princeps*) yet. Look for the mouse orthologue to find out the sequence identity between the mouse and human proteins, i.e. 80% (Target %id) and 82% (Target %id).

Selected orthologues

Show All entries | Show/hide columns | mouse

Species	Type	dN/dS	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Mouse (*Mus musculus*)	1-to-1	0.20465	ENSMUSG00000031138 F9 coagulation factor IX [Source:MGI Symbol;Acc:MGI:88384]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	X:59999464-60030759:1	80	82

The values are different due to the differences in length of the proteins (the mouse F9 protein is larger than its human counterpart). This is the alignment of the two proteins:



Orthologue alignment

Download homology

Type: 1-to-1 orthologues

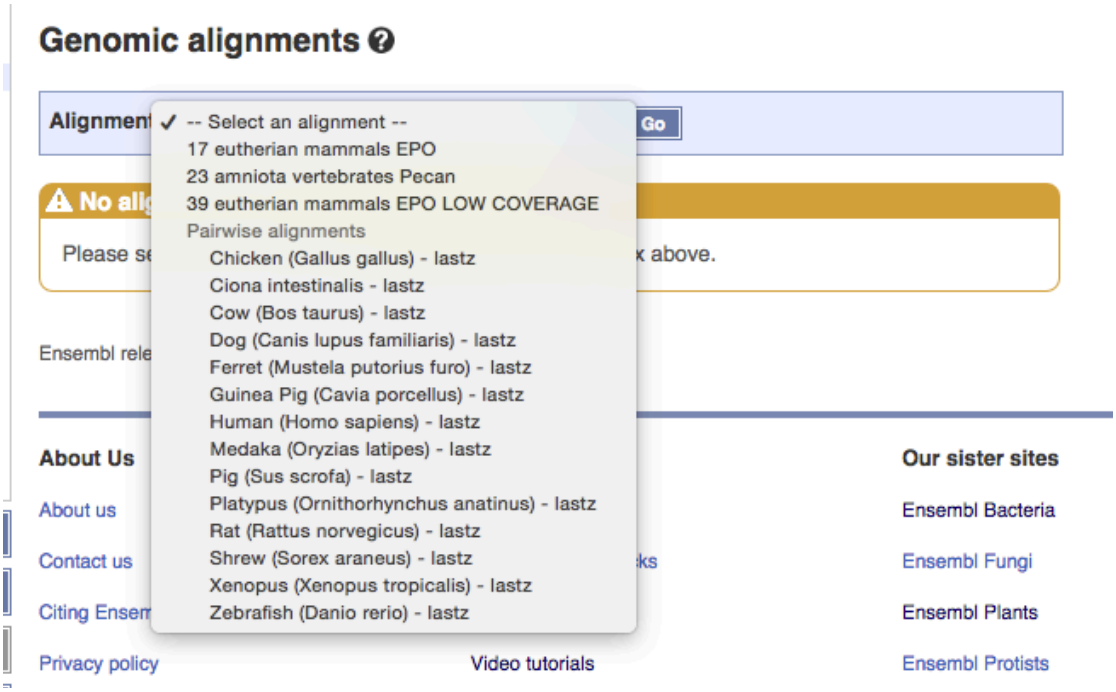
Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Human (*Homo sapiens*)	ENSG00000101981	ENSP00000218099	461 aa	82 %	100 %	X:139530758-139563458
Mouse (*Mus musculus*)	ENSMUSG00000031138	ENSMUSP00000033477	471 aa	80 %	98 %	X:59999464-60030759

```
CLUSTAL W(1.81) multiple sequence alignment


ENSP00000218099/1-461     MQRVNMIMAESPGLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSGKLEEFVQGNL
ENSMUSP00000033477/1-471  MKHLNTVMAESPALITIFLLGYLLSTECAVFLDRENATKILTRPKRYNSGKLEEFVRGNL
                          *:::* :*****.**** *******:**:****:***.***.***************:***


ENSP00000218099/1-461     ERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESNPCLNGGSCKDDINSYECWCP
ENSMUSP00000033477/1-471  ERECIEERCSFEEAREVFENTEKTTEFWKQYVDGDQCESNPCLNGGICKDDISSYECWCQ
                          ****:**:**************:********************** *****.******


ENSP00000218099/1-461     FGFEGKNCELDVTCNIKNGRCEQFCKNSADNKVVCSCTEGYRLAENQKSCEPAVPFPCGR
ENSMUSP00000033477/1-471  VGFEGRNCELDATCNIKNGRCKQFCKNSPDNKVICSCTEGYQLAEDQKSCEPTVPFPCGR
                          .****:*****.*********:******.****:*******:***:******:*******


ENSP00000218099/1-461     VSVSQTS-KLTRAETVFPDVDYVNST---------EAETILDNITQSTQSFNDFTRVVGG
ENSMUSP00000033477/1-471  ASISYSSKKITRAETVFSNMDYENSTEAVFIQDDITDGAILNNVTESSESLNDFTRVVGG
                          .*:* :* *:*******.::** ***            :**:*:*:*::*:*********
```

g) Click on the gene ID link of the mouse gene to jump to its page. You should be here:

http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000031138;r=X:59999464-60030759;t=ENSMUST00000033477

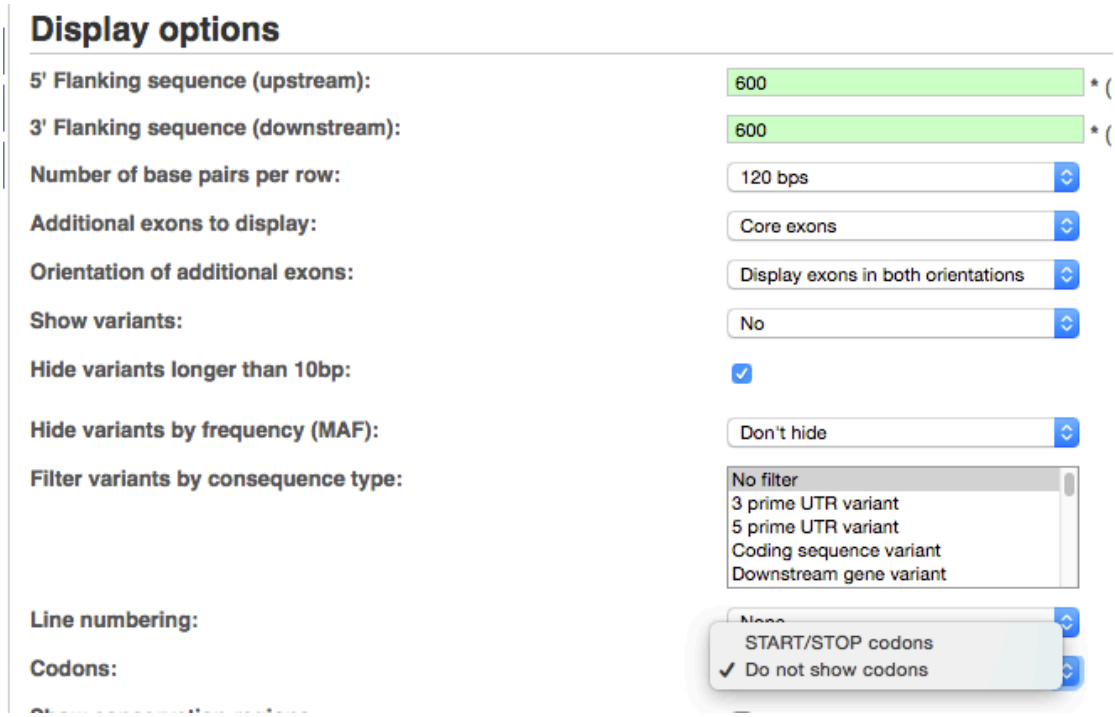
Now in the left hand menu, click on the 'Genomic alignments' link and choose human from the drop down menu:



You will see the alignment like the one below:

```
Mouse       GTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGTTTTTTGTTTGT
Human       ..................................................

Mouse       CTGGCTGTCCTGGAACTCACTTTGTAGACCAGGCTGGCCTCGAACTCA
Human       ..................................................

Mouse       TTATTATTAGATATTTTCTTTATTTACATTTCAAATGTTATCCCCTTT
Human       ..................................................

Mouse       CCTGGTCCTGGCATTCCCCTACACTGGGGCATTTGAGCCCTCACAGAA
Human       ..........................................CAGAA

Mouse       GAAATAGCCCAAAGATACACCGAGGGAGATGGACAACAATTTCCCAGA
Human       GAAATAGTCCAAAGACCCATTGAGGGAGATGGACATTA-TTTCCCAGA

Mouse       TCATGAAGCACCTGAACACCGTCATGGCAGAATCCCCGGCTCTCATCA
Human       TTATGCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCCTCATCA

Mouse       ATGCTTCTCCTTTAAAAACAGAAATATTTAGAACTACCTTTATCAAGT
Human       ATGCTTGCCTTTTAGATATAGAAATATCTGATGCTGTCTTCTTCAC-T
```

Click on the 'Configure this page' and select to show both START and STOP codons:



They will be highlighted in yellow:

Answers extra exercise 1 – Mouse assembly, protein domains

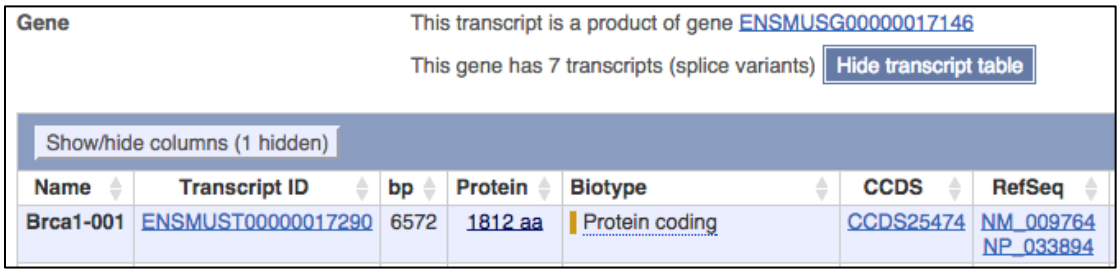
a) Go to the species page for mouse in Ensembl and click on 'More information and statistics'.

The current mouse genome assembly GRCm38 was released in January 2012. In release 83, Ensembl has annotated 12,923 no-coding genes including both short and long non-coding genes. The previous assembly of the mouse genome is also available in Ensembl (on our archive site) and corresponds to Ensembl release 67.

http://www.ensembl.org/Mus_musculus/Info/Index

b) The mouse Brca1 gene has got seven transcripts (splice variants, or alternatively spliced isoforms) and one of them has got a CCDS ID from the Consensus Coding Sequence project, a collaboration between EMBL-EBI, WTSI, NCBI and UCSC.

Go to the Trancript tab of ENSMUST00000017290 and click on 'Protein' under 'Sequence' at the top left for the amino acid sequence. Now click on the 'Protein Summary link to view the domains mapped to this protein, e.g. ZnF_RING and BRCT_domain. Alternatively click on '1812 aa' under the Protein column in the Transcript tab as shown below:



The domains are retrieved from various external databases, such as Superfamily, Pfam, among many others. Information on domains can be also displayed in a table format in 'Domains & features'. The table can be downloaded as a spreadsheet table.

Answers extra exercises 2 - Exploring a gene in Ensembl Bacteria

a) Enter part of the name of S. lividans 1326 strain into the genome search box (e.g. lividans 1326) and then select the correct genome to go to the info page. Enter 'era' into the search box and hit "Go". Click the "SLI_2875" next to Gene ID to go to the gene page of 'era'. From here, click "GO: molecular function" in the left hand menu. There are three terms GTPase activity, GTP binding and small ribosomal subunit rRNA binding corresponding to the accessions GO:0003924, GO:0005525 and GO:0070181

b) Go to the transcript tab and click on the protein summary link in the left hand side menu. The summary shows a GTP binding domain (TIGRFAM, Superfamily, Pfam, HAMAP) and a KH domain (Superfamily, Pfam, PROSITE profiles) though some predictions combine both domains into one signature.

Answers extra exercise 3 – miRNA genes in A. thaliana

a) Go to the Ensembl Plants homepage (http://plants.ensembl.org/). Select Arabidopsis thaliana and type mir395 in the search box.

There are six MIR395 family members in A. thaliana, i.e. MIR395A, -B, -C, -D, -E and -F.

b) The six MIR395 genes are organised in two clusters, one in the genomic region from base pair 9363196 to 9367179 on chromosome 1 (MIR395A, -B and –C) and one from base pair 26269979 to 26273969 on chromosome 1 (MIR395D, -E and –F).

Click on the genomic coordinates of the MIR395A gene, 1:9363196-9363288. The MIR395A gene is located on the reverse strand. Zoom out to view the other genes. MIR395B and -C genes are located on the forward strand (the second image in the Location tab is an overview image encompassing 200 kb. You can also see the orientation of the three genes in there. The signals > and < symbolise forward and reverse strand, respectively. Genes draw on the top of the blue bar (i.e. the genomic sequence) are on the forward strand, whereas reverse stranded genes are below the blue bar in the third image of the Location tab.

The MIR395D and -E genes are located on the reverse strand, while the -F gene is located on the forward strand.

The position of the genes relative to each other is identical within both clusters: two genes in opposite orientation to each other roughly 1 kb apart and the third gene roughly 2.5 kb away. Therefore, it has been hypothesised that the two clusters have evolved by tandem duplications, followed by an intrachromosomal duplication (Maher et al. Evolution of Arabidopsis microRNA families through duplication events. Genome Res 2006; 16:510-519).