Tutorial: Ensembl walkthrough

_____

In this tutorial you will walk through Ensembl using an example. You will explore the human **ABCD1** (ATP-binding cassette, sub-family D (ALD), member 1) gene.

The protein encoded by this gene is likely involved in the peroxisomal transport or catabolism of very long chain fatty acids (VLCFAs). Mutations in the *ABCD1* gene can cause **Adrenoleukodystrophy**, a rare X-linked disorder that causes a range of clinical phenotypes, often leading to a vegetative state and/or death (see also http://en.wikipedia.org/wiki/Adrenoleukodystrophy).

🖰 Go to the Ensembl homepage (http://www.ensembl.org/).

Searching

First of all, let's search for the human *ABCD1* gene.

🖰 Select 'Search: Human' and type 'abcd1' in the 'for' text box.
🖰 Click [Go].

The search result shows an *ABCD1* gene and several transcripts (splice variants).

🖰 Click on 'ABCD1 (Human Gene)' (the first hit)

This leads us to the 'Gene summary' page under the 'Gene' tab.

The Gene tab

Pages (also called 'views') in Ensembl are organised under a number of tabs, i.e. 'Species', 'Location', 'Gene', 'Transcript', 'Variation' and 'Regulation'. The various available pages under each tab are listed in the left-hand side menu.

The 'Gene Summary' page shows general information about the *ABCD1* gene and the transcripts that have been annotated for it as part of the GENCODE gene set (http://www.gencodegenes.org/).. Note the information icon (*i*) next to 'Gene summary' that opens up a help page, as well as the legend at the bottom of the graphical display.

🖰 Click [Show transcript table].

You can customise the table by clicking on 'Show/hide columns'. For example, turn on the UniProt matches, and turn off the Flags.

The graphical display (as depicted below) shows the same three transcripts as the table. You can click on a transcript to learn more about it.
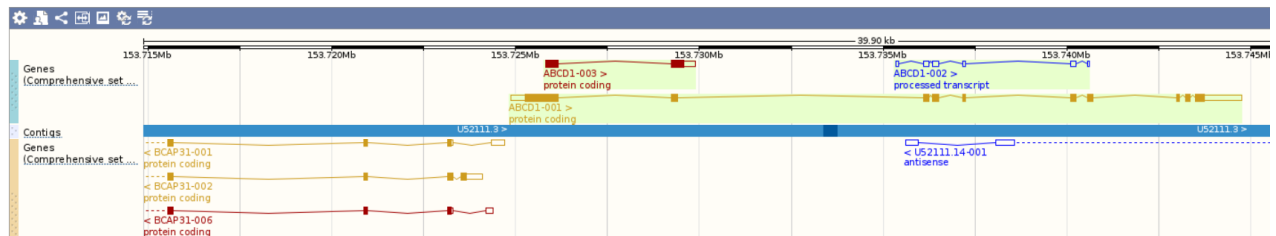
Protein-coding transcripts are gold or red. Gold transcripts are identical between the Ensembl and Havana projects, thus reflect a high standard. Red transcripts have either been annotated by Ensembl or Havana. In this case ABCD1-003 have been annotated by Havana. Non-protein coding transcripts are shown in blue.

Boxes and lines in the transcripts represent exons and introns, respectively. Empty boxes represent untranslated regions (UTRs), while filled boxes represent the coding sequence (CDS).

Summary – ABCD1 Transcripts

- There are three transcripts of which two are protein coding
- ABCD1-001 is gold, a symbol of high quality

The *ABCD1* gene is located on the forward strand of the genome. This can be seen from the arrows next to the transcript names, which indicate the direction of transcription and from the fact that the transcript models are shown above the blue bar that represents the genome. Transcripts located on the reverse strand are shown below the blue bar.

🖰 Click on 'Sequence' in the side menu.

Marked-up sequence ⓘ

[⬇ Download sequence] [✎ BLAST this sequence]

Key

Exons [ABCD1 exons] [All exons in this region]

```
>chromosome:GRCh38:X:153724268:153745362:1
CCTCGTCGATGGGCCGGGGAGCCTCCGCGGTCCCGGAGCCCAGCCCGGCGCGCGGAGCCC
GCTCACCGAGTTTCCCACAGTCAACGTGCAGGCCCCGCCGCAGCAACAGAACTCTCCCAC
AGCAGCCCCGGCCCCGCCCCTCATACCGCGGCCGGAAACCGGAAGCGCCCGCCGGGCACC
GCCCACCAGCCCTCGCGAGGCCCCGGAGGCTCCGCCCACCTCCGCTTCCCACCCCGCCCC
GGAGCGGAGGGCCGGCGCTCCGAGCGGGAGAGGAAGAGGCGCCTCGGGCTCCGGGCGAGC
AGGGCGGGGTGGAGCGAGCACGCGGGCGGGGCGGGGCGGGGCTTTGTCGGGCCGGCGAGG
GCCGCTTCTCTAGTCCGCGCGGCCGTCCCACGTCTCTGTGGTGCGGGAGGGGCCCCGCCG
AGGGGCGAGAACGGGAGGTGGGGGGTGTGGGCGGGCCCCGCCGAGGGGCGAGAACAGGGTG
GGGCTCCCGCGCCCGGACTCCGCCCCTCGCCCCTCCTCCGCCTCCTCCCCTTCCCCCGAC
TCGCCCCTGGGGAAGAGTGGGTGGGGATTCTGGGCCGGTGGAGGAGTCACTGTCGCTTCA
GCCAGGCTGCGGAGCGGACGGACGCGCCTGGTGCCCCGGGGAGGGGCGCCACCGGGGGAG
GAGGAGGAGGAGAAGGTGGAGAGGAAGAGACGCCCCCTCTGCCCGAGACCTCTCAAGGCC
CTGACCTCAGGGGCCAGGGCACTGACAGGACAGGAGAGCCAAGTTCCTCCACTTGGGCTG
CCCGAAGAGGCCGCGACCCTGGAGGGCCCTGAGCCCACCGCACCAGGGGCCCCAGCACCA
CCCCGGGGGCCTAAAGCGACAGTCTCAGGGGCCATCGCAAGGTTTCCAGTTGCCTAGACA
ACAGGCCCAGGGTCAGAGCAACAATCCTTCCAGCCACCTGCCTCAACTGCTGCCCCAGGC
ACCAGCCCCAGTCCCTACGCGGCAGCCAGCCCAGGTGACATGCCGGTGCTCTCCAGGCCC
CGGCCCTGGCGGGGGAACACGCTGAAGCGCACGGCCGTGCTCCTGGCCCTCGCGGCCTAT
GGAGCCCACAAAGTCTACCCCTTGGTGCGCCAGTGCCTGGCCCCGGCCAGGGGTCTTCAG
```

→ **Exon from neighbouring gene (BCAP31)**

→ **ABCD1 Exon**

On the 'Sequence' page the sequence of the *ABCD1* gene plus 600 bp upstream and downstream is shown. Exon sequences belonging to the *ABCD1* gene are shown in red letters on a peach background, while exons belonging to other genes are shown in black letters on a peach background. All possible exon sequence is shown, across all the transcripts.

Almost all graphical displays in Ensembl can be configured. This is done using the [Configure this page] button.

🖑 Click [Configure this page] in the side menu.

A pop-up window lists all display options.

Summary – Gene Sequence

- All exons are highlighted, for all genes in the region
- ABCD1 exons are in bold letters

Let's find out more about *ABCD1.*

🖰 Click on External References in the side menu.

This shows matches to the Ensembl gene in other projects and databases. A table that links Ensembl transcripts to UniProt and RefSeq identifiers is found at the bottom of the page.

The following database identifiers correspond to the transcripts of this gene:

Transcript ID	CCDS	UniProtKB/ Swiss-Prot	RefSeq peptide	RefSeq mRNA	Vega transcript	UniProtKB/ TrEMBL
ENST00000218104	CCDS14728.1	P33897	NP_000024.2	NM_000033.3	OTTHUMT00000061041	
ENST00000370129					OTTHUMT00000061043	A6NEP8
ENST00000443684					OTTHUMT00000061042	

This is similar to what we saw in the Transcript table, but has more information.

You can again customise the table by clicking on 'Show/hide columns'.   For example, turn on Gene ontology (GO) classes and hide the RefSeq IDs

The following database identifiers correspond to the transcripts of this gene:

Transcript ID	UniProtKB/ Swiss-Prot	UniProtKB/ TrEMBL	GO
ENST00000218104.5	P33897		GO:0000166 GO:0005215 GO:0005324 GO:0005324 GO:0005325 GO:0005325 GO:0005515 GO:0005515 GO:0005524 GO:0005524 GO:0005737 GO:0005739 GO:0005777 GO:0005777 GO:0005778 GO:0005778 GO:0005778 GO:0005778 GO:0005778 GO:0005779 GO:0005829 GO:0006635 GO:0006635 GO:0006635 GO:0006810 GO:0006810 GO:0007031 GO:0015910 GO:0015910 GO:0015919 GO:0016020 GO:0016021 GO:0016021 GO:0016887 GO:0016887 GO:0019899 GO:0033540 GO:0036109 GO:0042626 GO:0042626 GO:0042758 GO:0042760 GO:0042803 GO:0043651 GO:0048471 GO:0055085 GO:0055085
ENST00000370129.4		A6NEP8	GO:0005524 GO:0005778 GO:0006635 GO:0006810 GO:0016021 GO:0042626 GO:0055085
ENST00000443684.2			

🖰 Click on 'Phenotypes' in the side menu.

On the 'Phenotypes' page phenotypes that have been associated with the *ABCD1* gene as well as with variants associated with the *ABCD1* gene are shown.

Phenotypes ❓

Phenotype(s), disease(s) and trait(s) associated with this gene ENSG00000101986

Phenotype, disease and trait	Source	Allelic requirement	Genomic locations
ADRENOLEUKODYSTROPHY	MIM morbid	-	View on Karyotype
Adrenoleukodystrophy, X-Linked	DDG2P	hemizygous	View on Karyotype
ADRENOMYELONEUROPATHY	Orphanet	-	View on Karyotype
CADDS	Orphanet	-	View on Karyotype
X-linked cerebral adrenoleukodystrophy	Orphanet	-	View on Karyotype

Phenotype, disease and trait annotations associated with variants in this gene

Phenotype, disease and trait	Source(s)	Genomic locations	Biomart	Number of variants	Show/hide details
ALL variants with a phenotype annotation	-			642	Show
ADRENOLEUKODYSTROPHY	ClinVar, OMIM, Uniprot	View on Karyotype	-	55	Show
ADRENOMYELONEUROPATHY	OMIM	View on Karyotype	-	5	Show
Addison disease, susceptibility to	OMIM	View on Karyotype	-	1	Show
Addisons disease	ClinVar	View on Karyotype	-	1	Show
Annotated by HGMD but no phenotype description is publicly available	HGMD-PUBLIC	-	-	569	Show
ClinVar: phenotype not specified	ClinVar	View on Karyotype	-	20	Show

🖱 Click on 'GO: Biological process' in the side menu.

Gene Ontology (GO) terms (http://www.geneontology.org) associate proteins to biological process, molecular function and cellular component terms.

GO: Biological process ❓

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0006635	fatty acid beta-oxidation	IEA	InterPro	ENST00000370129 ENST00000218104	• Search BioMart • View on karyotype
GO:0006810	transport	IEA	InterPro	ENST00000370129 ENST00000218104	• Search BioMart • View on karyotype
GO:0007031	peroxisome organization	IDA, NAS	UniProt	ENST00000218104	• Search BioMart • View on karyotype
GO:0015910	peroxisomal long-chain fatty acid import	IEA	InterPro	ENST00000218104	• Search BioMart • View on karyotype
GO:0015919	peroxisomal membrane transport	NAS	UniProt	ENST00000218104	• Search BioMart • View on karyotype
GO:0033540	fatty acid beta-oxidation using acyl-CoA oxidase	TAS	Reactome	ENST00000218104	• Search BioMart • View on karyotype
GO:0036109	alpha-linolenic acid metabolic process	TAS	Reactome	ENST00000218104	• Search BioMart • View on karyotype

The 'biological process' terms indicate that the ABCD1 protein plays a role in fatty acid transport and catabolism.

🖱 Click on 'GO: Cellular component' in the side menu.

The 'cellular component' terms indicate the ABCD1 protein is located in the peroxisomal membrane.

GO: Cellular component ❓

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0005737	cytoplasm	IDA	UniProt	ENST00000218104	• Search BioMart • View on karyotype
GO:0005739	mitochondrion	IEA	Ensembl	ENST00000218104	• Search BioMart • View on karyotype
GO:0005777	peroxisome	IEA	InterPro	ENST00000218104	• Search BioMart • View on karyotype
GO:0005778	peroxisomal membrane	IEA	InterPro	ENST00000370129 ENST00000218104	• Search BioMart • View on karyotype
GO:0005779	integral component of peroxisomal membrane	IDA, NAS	UniProt	ENST00000218104	• Search BioMart • View on karyotype
GO:0005829	cytosol	TAS	Reactome	ENST00000218104	• Search BioMart • View on karyotype
GO:0016020	membrane	IDA, IEA	UniProt	ENST00000218104	• Search BioMart • View on karyotype
GO:0016021	integral component of membrane	IEA	InterPro	ENST00000370129 ENST00000218104	• Search BioMart • View on karyotype
GO:0048471	perinuclear region of cytoplasm	IDA	UniProt	ENST00000218104	• Search BioMart • View on karyotype

🖰 Click on 'ENST00000218104.5' in the transcript table at the top of the page.

This leads us to the 'Transcript summary' page under the 'Transcript' tab.

The Transcript tab

Note that, because we have moved from the 'Gene' tab to the 'Transcript' tab, the side menu has changed and now shows links to pages with information about this specific splice variant.

🖰 Click on 'Sequence - Exons' in the side menu.

On the 'Exons' page the sequence of the unspliced transcript is shown. The translated region is shown in blue, the flanking sequences in green, introns in grey and untranslated regions in red. By default only a small part of the introns and the flanking sequences is shown, but this can be changed on the configuration page.

🖰 Click on 'External References - General identifiers' in the side menu.

On the 'General identifiers' page cross-references to other databases are shown that contain entries that correspond to the ENST00000218104.5 sequence.

This transcript corresponds to the following database identifiers:

Show All entries

External database	Database identifier
CCDS	CCDS14728.1 [view all locations]
Eukaryotic Promoter Database (Bucher)	P33897 [view all locations]
European Nucleotide Archive	BC015541 [align] [view all locations] BC025358 [align] [view all locations] U52111 [align] [view all locations] Z21876 [align] [view all locations] Z31006 [align] [view all locations] Z31007 [align] [view all locations] Z31008 [align] [view all locations] Z31009 [align] [view all locations] Z31010 [align] [view all locations] Z31348 [align] [view all locations]
HGNC transcript name	ABCD1-001 ATP binding cassette subfamily D member 1 [view all locations]
Havana translation	OTTHUMP00000025960 [view all locations]
INSDC protein ID	AAH15541.1 [align] [view all locations] AAH25358.1 [align] [view all locations] CAA79922.1 [align] [view all locations] CAA83230.1 [align] [view all locations]
Reactome	R-HSA-1369062 ABC transporters in lipid homeostasis [view all locations] R-HSA-1430728 Metabolism [view all locations] R-HSA-1643685 Disease [view all locations] R-HSA-2046104 alpha-linolenic (omega3) and linoleic (omega6) acid metabolism [view all locations] R-HSA-2046105 Linoleic acid (LA) metabolism [view all locations] R-HSA-2046106 alpha-linolenic acid (ALA) metabolism [view all locations] R-HSA-382551 Transmembrane transport of small molecules [view all locations] R-HSA-382556 ABC-family proteins mediated transport [view all locations] R-HSA-390247 Beta-oxidation of very long chain fatty acids [view all locations] R-HSA-390918 Peroxisomal lipid metabolism [view all locations] R-HSA-556833 Metabolism of lipids and lipoproteins [view all locations] R-HSA-5619084 ABC transporter disorders [view all locations] R-HSA-5619115 Disorders of transmembrane transporters [view all locations] R-HSA-5684045 Defective ABCD1 causes adrenoleukodystrophy (ALD) [view all locations]
RefSeq mRNA	NM_000033.3 [Target %id: 100; Query %id: 99] [align] [view all locations]
RefSeq peptide	NP_000024 [Target %id: 100; Query %id: 100] [align] [view all locations]
UCSC Stable ID	uc004fff.2 [view all locations]
UniParc	UPI0000000DF5 [view all locations]
UniProtKB/Swiss-Prot	P33897 [Target %id: 100; Query %id: 100] [align] ATP-binding cassette sub-family D member 1 [view all locations]

For example, ENST00000218104 matches the P33897 protein sequence in the UniProtKB/Swiss-Prot database (http://www.uniprot.org/) and the NP_000024.2 protein and NM_000033.3 mRNA sequence in the RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/).

Let's have a look at the region on the chromosome, and neighbouring genes.
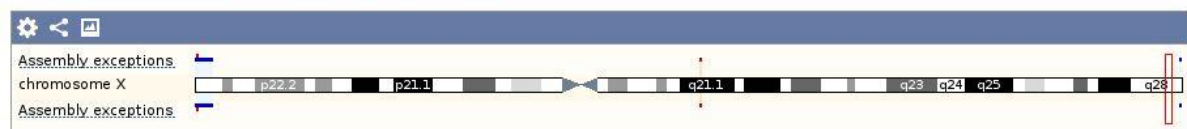
🖱 Click on the 'Location' tab.

This leads us to the 'Region in detail' page under the 'Location' tab.

The Location tab

The 'Region in detail' page shows the genomic neighbourhood of the *ABCD1* gene. It consists of three parts.
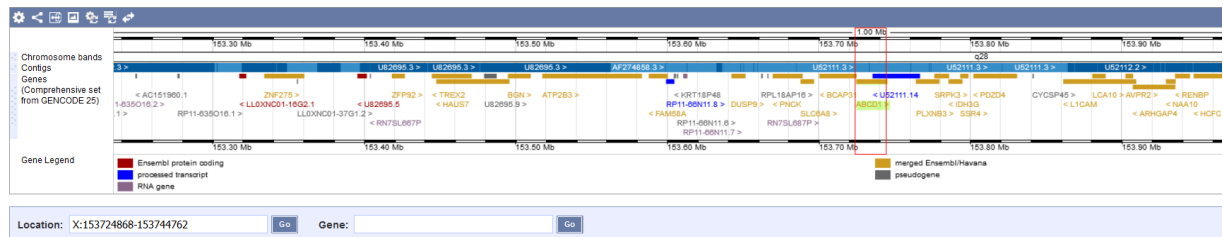
First, the complete chromosome.

Chromosome X: 153,724,868-153,744,762

Second, the 1 Mb region around the gene of interest.

Region in detail

Drag down the bottom to reveal *ABCD1*.

This display is scrollable. Either use the 'Scroll' arrows click and drag the image in the same way as Google Maps. Zoom in by clicking the 'Drag/Select' icon, selecting the region of interest with your mouse and subsequently clicking 'Jump to region' in the resulting pop-up.

Third, the region of interest. In our case this is the *ABCD1* gene.
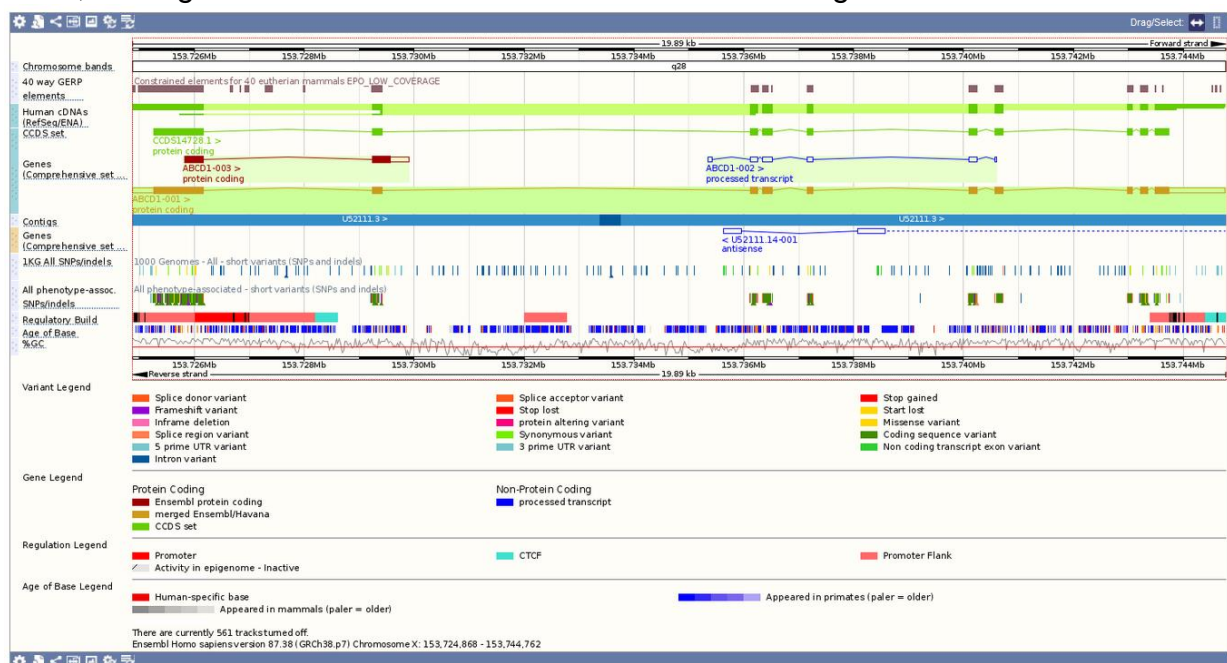
By default, the data tracks drawn are:

- 40 way GERP elements (the 'constrained elements', which are regions of high conservation based on comparison of sequence across 40 species)
- Human cDNAs (cDNA sequences aligned to the genome)
- CCDS set (transcripts in the Consensus Coding Sequence Set)
- Genes (GENCODE)
- Contigs (the genome)
- 1KG-All-common (Variants from the 1000 Genomes project with population frequency >1%)
- All phenotype-associated variants
- 1000 Genomes High Quality Structural Variants
- MultiCell regulatory features (sequences that may be involved in gene regulation)
- %GC (reflects GC content vs AT)

There are several ways to navigate this display:

- zoom in and out by using the [+/-] slider
- at Drag/Select you can zoom in by drawing a box around the region of interest and subsequently clicking 'Jump to region' in the resulting pop-up or by clicking at the ↔ sign to move the window.
- going to a particular region by changing the coordinates in the 'Location' text box or by searching for a gene using the 'Gene' text box (which has auto completion)

Datasets (or tracks) can be added to the display using [Configure this page]. On the configuration page all available tracks are grouped in the left-hand menu. It is also possible to search for tracks using the 'Find a track' text box.

For example, to add protein alignments from UniProt to the display:

🖰 Click [Configure this page] in the side menu.
🖰 Type 'UniProt in the 'Find a track' text box.
🖰 Select 'Proteins (mammal) from UniProt'. Choose 'Normal'.
🖰 Click (✓).
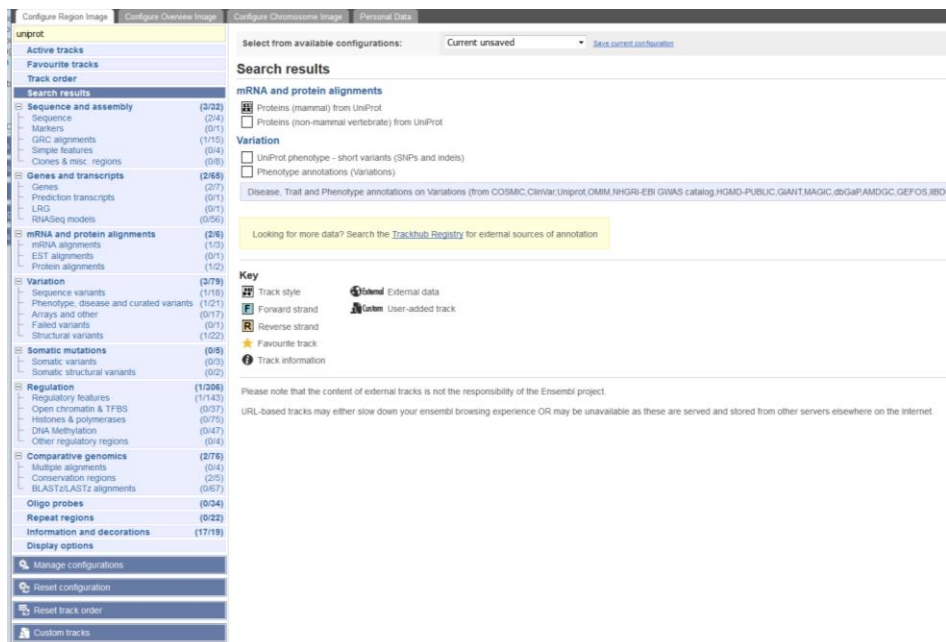
A new track, 'UniProt (mammals)', has now been added to the display.
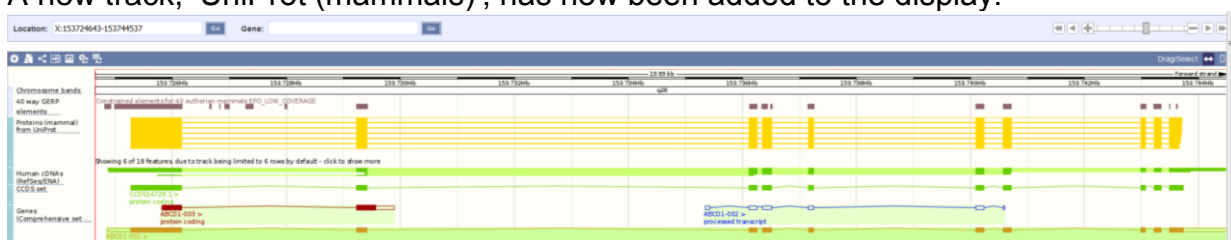
To turn the added track off again:

🖱 Hover over the track name.
🖱 Click on the 'Turn track off' icon (x) in the pop-up.

Tracks can be moved by clicking on the bar in front of the track name and dragging the track to the desired location.

To zoom in, you can click and drag your mouse around a region.

🖱 Zoom in to ABCD1-002

At the top of the display (circled in the image above) several icons are shown, some of which can also be found on other displays:

- Configure this image: add/delete tracks (same as [Configure this page] button in the side menu). ⚙

- Manage your custom tracks: add your own data (same as [Add your data] button in the side menu) 📑

- Share this image: create a URL that can be shared with others without the need to tell them how to configure the page ◁

- Resize this image: resize the image ⊞

- Export this image: export the image in various formats (PDF, PNG etc.) 🖼