7. Bioinformatics practice

### [Ensembl genome browser 94](https://www.ensembl.org/)

Open a new Word document and please write / paste your answers to this document! (Or you can write your answers to this document as well. If you push Shitf+Enter combination at the end of the lines, the number of the lines will not change.)

Create a folder, called practice\_7 and save your work to this directory

1. Open in the browser https://www.ensembl.org/index.html

### Write down some information about the newest genome of the cattle! All genomes -> Select species cow

How many base pair the assembly have? More information and statistics->Statistics /Summary

How many coding genes are there? More information and statistics->Statistics/Gene counts

### How many chromosomes it have? View karyotype

1. Search the merged Ensembl/Havana genes that are in the region (55169750-55388934) of the 16th chromose of the newest human assembly! What genes could you find? Write down their names (HGNC) and numbers.

Search-> human for chr16:55,169,750-55,388,934 GO

1. Search your protein’s gene in the human genome. (In the case of ERK1 protein, search with MAPK3 instead)

Where it is? (gene name Human Gene) chr:start-end: orientation (Which strand is located on?)

How many transcripts it has? About this gene or click on the gene name-> Show transcript table

1. How many transcripts the first coding transcript has, and how many exons which is coding? How long the transcript (base pair) and the protein that came from this (amino acid)?

Click on the first transcript in the table the in the Transcript ID column.

On the image the boxes mean exons and the lines which connect the exons are introns. If the boxes are colorful, the sequences are coding sequences, if not the sequences are translated regions. The color of the transcripts means the biological type: yellow is the automatic Ensembl and annote with the annotation pipeline and furhtermore the transcript is merged with the genes identified in Havana (good quality). The red one shows the automaticly annotated transcript of the Ensembl. The blue ones are not coding.

1. What is the Ensembl identifier of the 1. transcript, which is coding a protein and what is the identifiers of the protein and the gene? Is there any SNP in the first exon of the first transcript which can result a stop codon? If there any, what is its name and what base or amino acid change happened?

ENS + (ENSEMBL) + MUS (3 letters means the species, there isn’t of the human!) + G/T/P/E/R (gene/transcript/protein/exon,/regulation information) + 11 number

Click on the first transcript then on the left side Summary -> Exons,

According to the table search a stop variant in the sequence of the 1. ENSE (11 number). Click on it and get the informations from the appearing window.



1. Search for the first frameshift mutation, if there is any! What the name of its SNP and what was the ancient nucleotide? Search for the first purple nucleotide, click on it and the appearing window click on the Explore this variant option.
2. Go back to your gene (Gene: your gene name). On the left side you can find the Sequence option. Choose the BLAST this sequence button under the Marked-up sequence. At the Search against option, choose the Ad remove your protein option. To the Species selector, type: Chimpanzee. Click on the Apply button. With DNA sequence, search for DNA sequence! In the search tool, choose the BLASTN program. Run the program with the Run button. In which chimpanzee’s chromose, did you get a found?

8) Get some information about the chromosome region with BioMart

1. Setup the filter of the BioMart to that region (arm, number) according to your protein location (Filters, Region, Chromosome/scaffold and Karyotype band (start, stop are the same)).

How many genes are in this region? (Count button)?

1. You need the further informations about the (Attributes):

(GENE) Gene name (Gene Name), Ensembl identifier (Gene stable ID)

the phenotype, disease (EXTERNAL: MIM Morbid Description)

Save the result (Results) unique (Unique results only) table in tsv format and send to yourself via e-mail (Compressed web file notify by email).

Is there any phenotype or disease that is characteristic of several genes, proteins?

+2 tasks (does not compulsory)

9) According to your gene which primate mammal is closest to humans? Make a tree! (Gene-based displays-[Comparative Genomics](https://www.ensembl.org/Homo_sapiens/Gene/Compara?db=core;g=ENSG00000175387;r=18:47808957-47931146)- [Gene tree](https://www.ensembl.org/Homo_sapiens/Gene/SpeciesTree?db=core;g=ENSG00000175387;r=18:47808957-47931146) and [Gene gain/loss tree](https://www.ensembl.org/Homo_sapiens/Gene/SpeciesTree?db=core;g=ENSG00000175387;r=18:47808957-47931146) )

10) What kind of RNA-seq sequences are in your gene region? In which tissue is expressed to a large extent from this five (brain, fat, blood, kidney, kidney)? (Configure this page -RNA-seq models, see in the tutorials)