Practicing Ensembl

Ensembl genome database project is a scientific project at the European Bioinformatics Institute, which was launched in 1999 in response to the imminent completion of the Human Genome Project. It is a ***genome browser*** for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. It interprets genes, computes multiple alignments, predicts regulatory function and collects disease data. **Ensembl** tools include **BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP)** for all supported species.

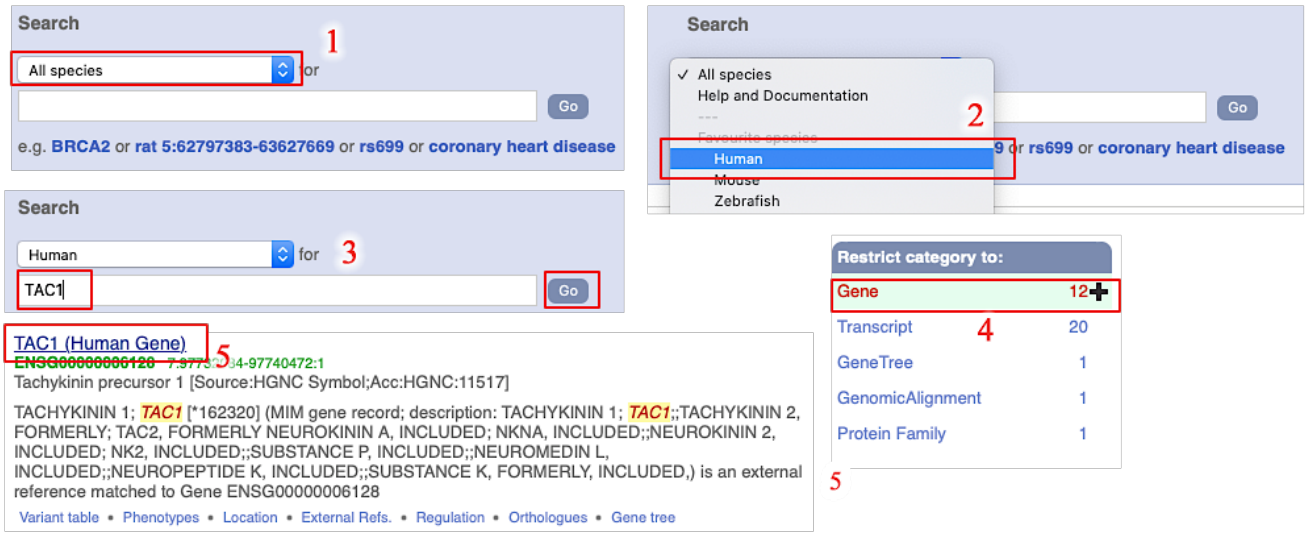
To exhibit how to use and search Ensembl, a search on the human TAC1 gene was conducted for further guidance throughout the website.

- *Tachykinins (TAC) are active peptides which excite neurons, evoke behavioral responses, are potent vasodilators and secretagogues, and contract (directly or indirectly) many smooth muscles.*

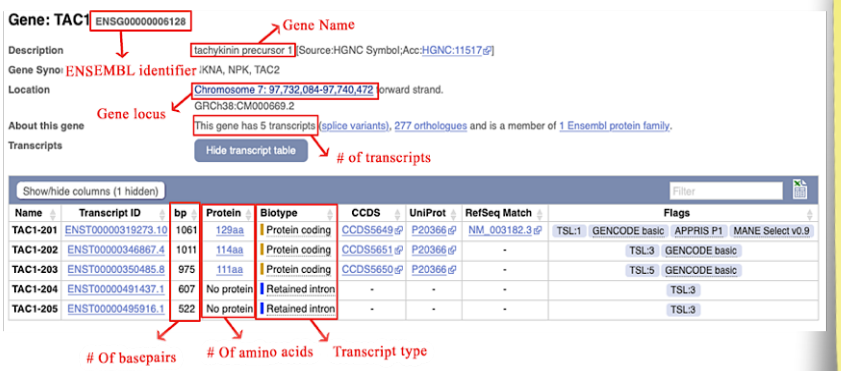
- ***How to search for a gene in ENSEMBL***

1- **Conduct your search by determining the species, in this example → human**
2- **Write the gene of interest, in this example → TAC1**
3- **Restrict the search to Gene**
4- **Select TAC1 (Human Gene)**

Press Go



- ***How to explain your search results.***



Note:

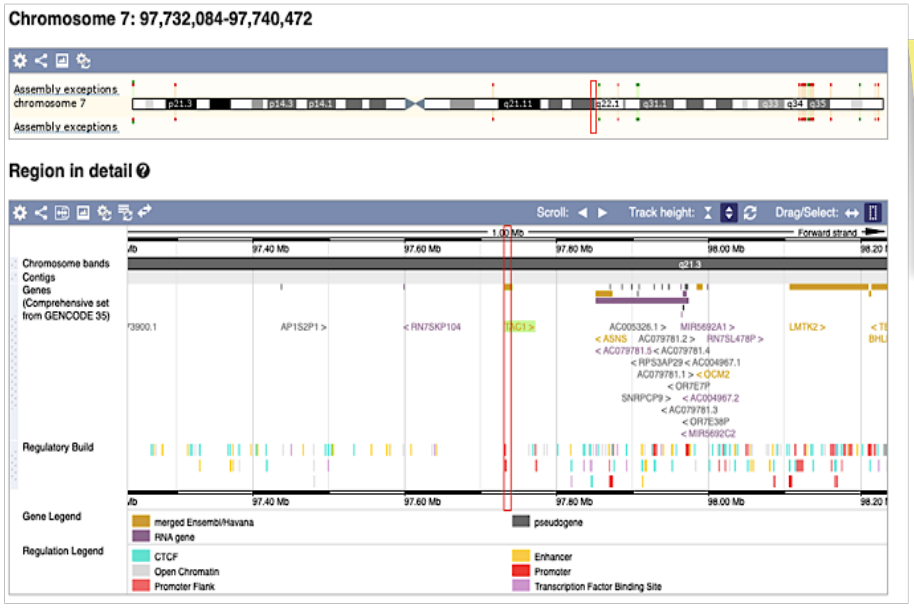
- Ensembl identifier = ENS + T → Transcript + number
  G → gene
  P → protein

- *Splice variants* are more than one transcript of a gene due to alternative splicing. Therefor they differ in the number of base pairs and amino acids.

- ***How to identify gene location in details.***

e.g.1 *On which chromosome is this gene located? Show the graphical position of the gene on the chromosome (Region in details).*

e.g.2 *Is the gene transcribed from the forward or from the reverse strand of the genome assembly?*



Note:

To save a picture of the graphical position and regions, export the image as a pdf or presentation file

- ***How to retrieve information about the gene's transcripts.***

Information of transcripts are displayed in the transcript table

Transcripts

Hide transcript table

Show/hide columns (1 hidden)

Name	Transcript ID	bp	Protein	Biotype	CCDS
TAC1-201	ENST00000319273.10	1061	129aa	Protein coding	CCDS5649
TAC1-202	ENST00000346867.4	1011	114aa	Protein coding	CCDS5651
TAC1-203	ENST00000350485.8	975	111aa	Protein coding	CCDS5650
TAC1-204	ENST00000491437.1	607	No protein	Retained intron	-
TAC1-205	ENST00000495916.1	522	No protein	Retained intron	-

e.g.3 *How many transcripts (splice variants) has Ensembl annotated for it?*

It has 5 transcripts, 3 of which are protein coding, while 2 transcripts have retained intron.

e.g.4 *What is the longest transcript, and how long is the protein it encodes?*

The longest transcript is TAC1-201, it has 1061 bp and it encodes for a 129aa protein.

e.g.5 *Which transcript has a CCDS record associated with it?*

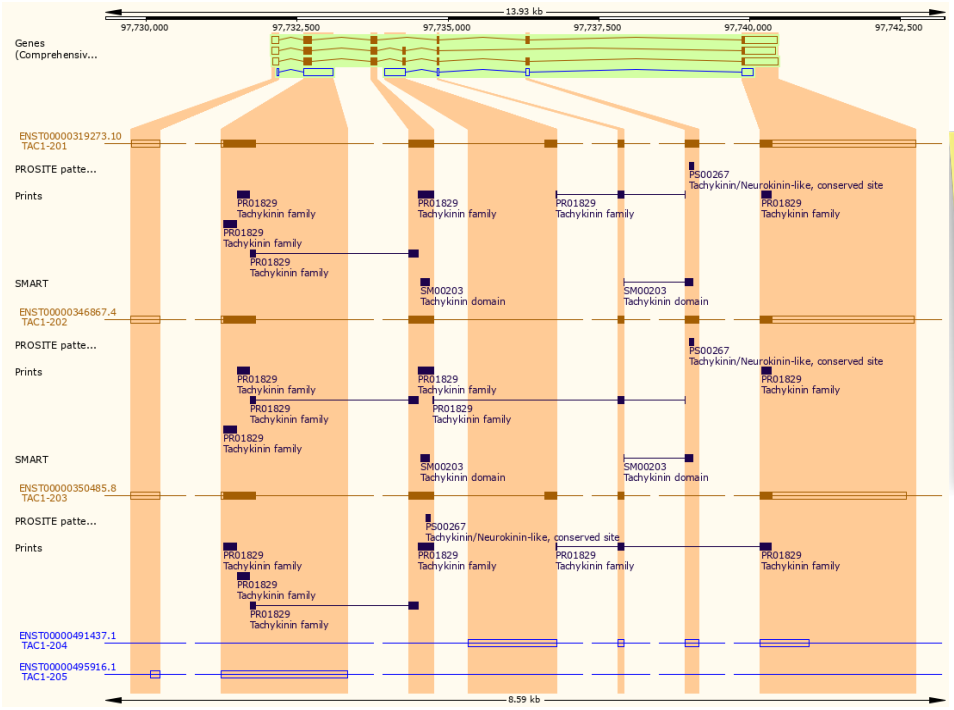
TAC1-201, TAC1-202, and TAC1-203

Note:

- Consensus coding domain sequence (CCDS) is an ID number for protein coding transcripts that provides an overall information about the gene and its proteins.

To show how each variant differ from the other (Transcript Comparison) by showing the structure (exons) for each one. → *press the hyperlink titled (splice variants), a graphical view of every transcript will appear.*

About this gene This gene has 5 transcripts (splice variants), 277 orthologues and is a member of 1 Ensembl protein family.

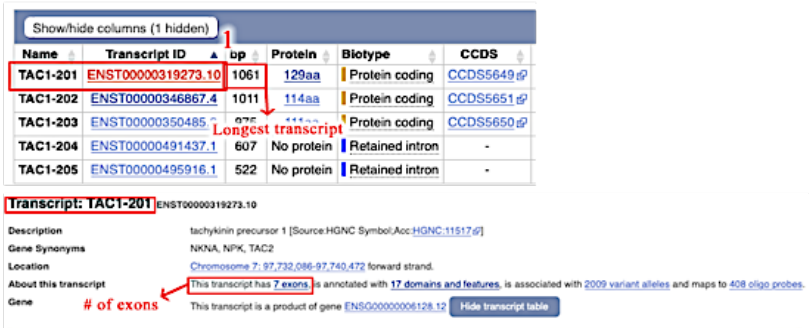


Note:

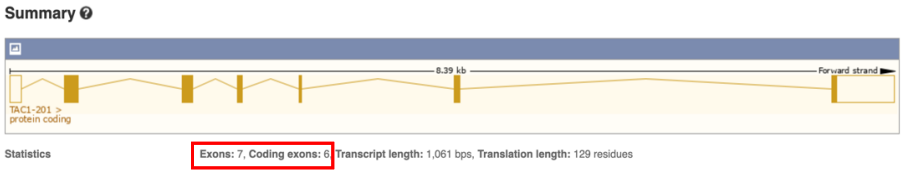
- Transcripts are drawn as boxes (exons) and lines connecting the boxes (introns).
- Filled boxes represent coding sequence and unfilled boxes (or portions of boxes) represent untranslated regions (UTR)".

- ***Gene exons explained in details***

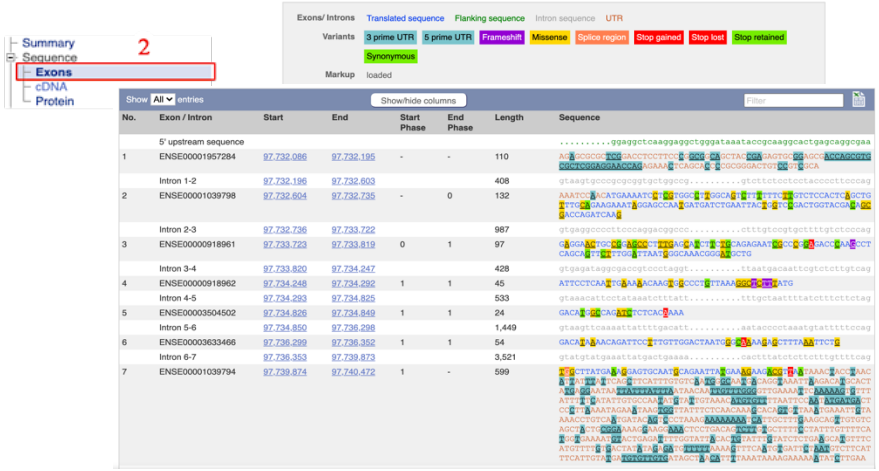
e.g.6 *How many exons does the longest transcript have?*



e.g.7 Are any of its exons completely or partially translated?



More info about the translated regions of exons and their sequence are provided in the side bar summary hierarchy under ***sequence***.



Note:

- *To identify which exons are completely or partially translated, check the colour coding of the sequence of each exon*
  - *Fully blue → completely translated*
  - *Fully red → completely untranslated*
  - *A mix of both colours → partially translated*

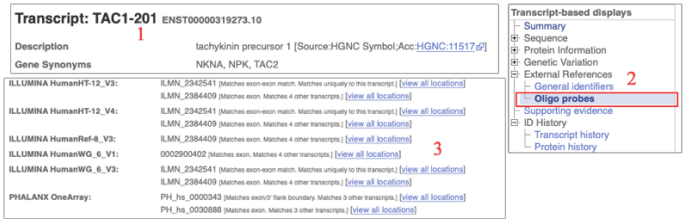
→ **Therefor, the answer is 2 exons** are partially translated, **4 exons** are completely translated, **1 exon** is untranslated

- ***Microarray probe sets available for gene expression***

Microarrays are used to measure the expression levels of large numbers of genes simultaneously. One of the applications that is provided by Ensembl is to annotate expression microarrays on the reference genome and transcripts sequences for those arrays whose manufacturers disclose the probe sequences. This probe is a short DNA sequence targeting a short region of a transcript. They are used to detect the presence of nucleotide sequences through hybridization to single-stranded nucleic acid due to complementarity between the probe and the target.

e.g8 *Is it possible to monitor the expression of TAC1-201 using the Illumina microarray?*

Yes, it is possible. This information can be obtained from the side bar summary hierarchy under ***external references*** → oligo probes. These probes are identified with an ID number (ILMN_no.) that can be ordered from the manufacturing company.

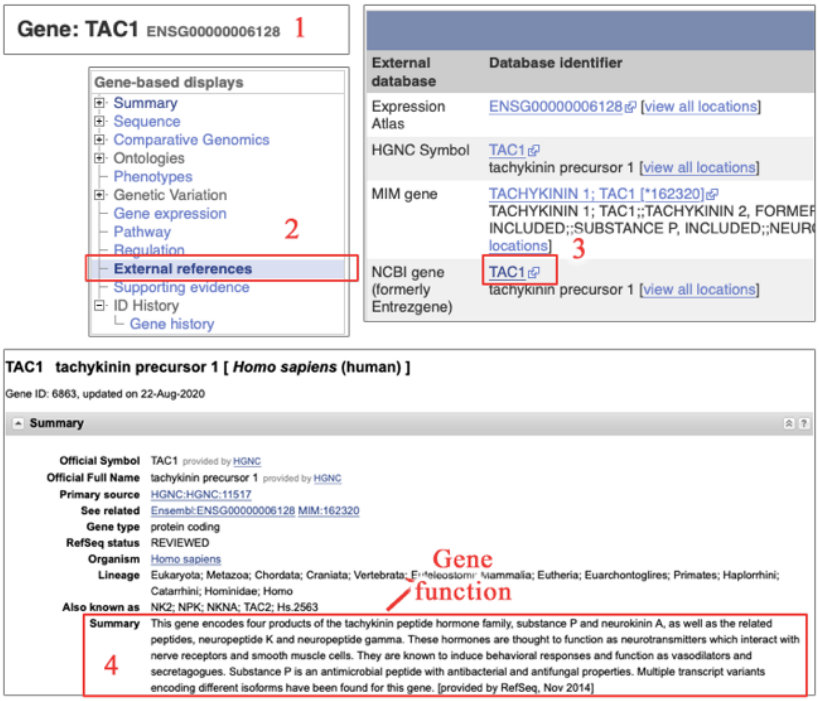


- ***How to retrieve the function of gene using External References***

e.g.9 *Have a look at the External References. What is the function of TAC1?*

Make sure to check (Gene tab) at the top of the page

External references is found at the sidebar → to check the function → click the hyperlink in NCBI gene

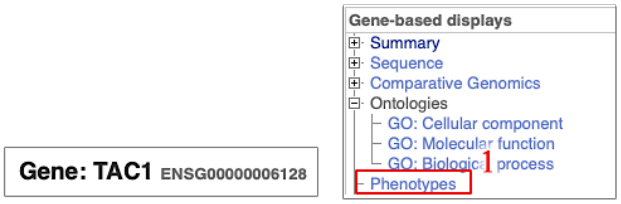


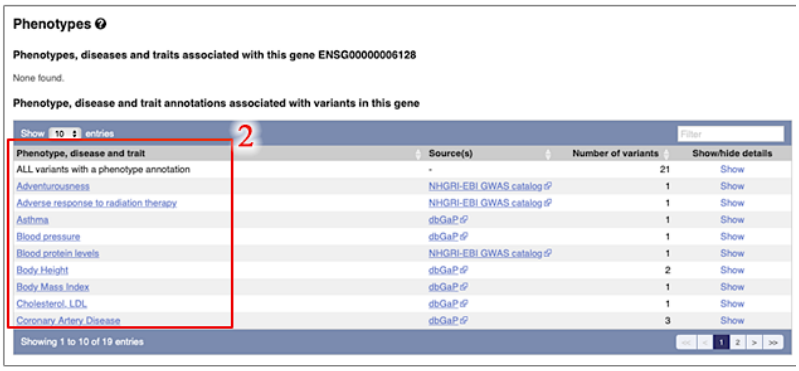
- ***Diseases associated with a gene***

e.g.10 *Are there any diseases associated with variants in this gene?*

Make sure to check (Gene tab) at the top of the page

Phenotype, disease and trait annotations associated with variants in this gene is found at ***the sidebar →Phenotypes***



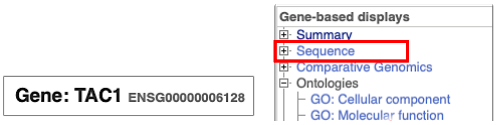


- ***How to retrieve a gene DNA sequence***

e.g.11 *Retrieve the TAC1 gene sequence.*

Make sure to check (Gene tab) at the top of the page

Gene sequence can be retrieved from ***the sidebar →Sequence***



Marked-up sequence

Download sequence | BLAST this sequence

Exons TAC1 exons All exons in this region

Markup loaded

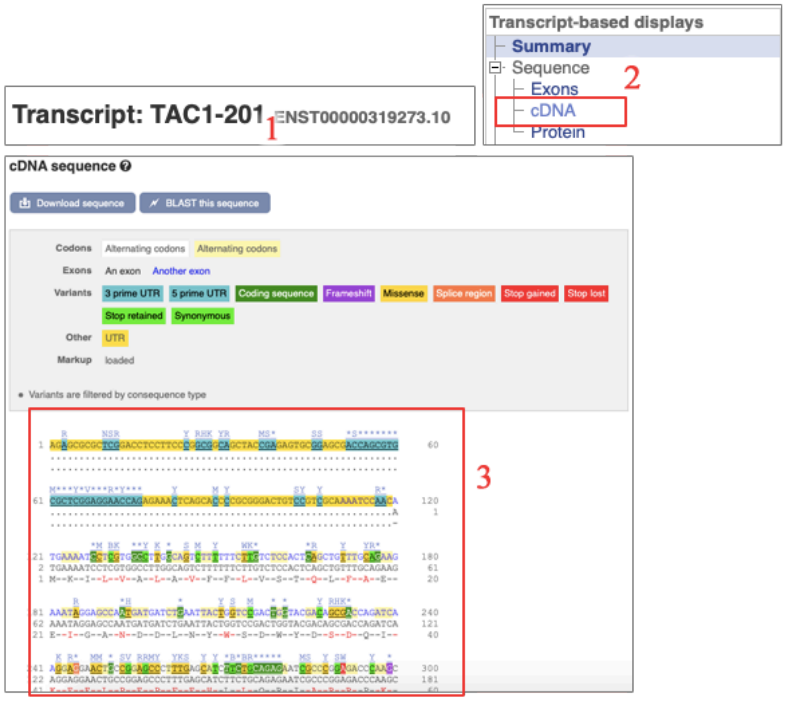
```
>chromosome:GRCh38:7:97731484:97741072:1
AGGAAAAGCCAGTATTCGCGTTGATTTAGAAGAGGGATGTTCTGGTTATAGAACGATGCT
GTGTCTCAGAAACACTTAAATACTATTAAGCTAGAAATAGAAGGGAAAATAATGCTTCCC
CGCATCTCCCCTCAAGTGTAGTCCTCTTTTTTTAGCCTGATTTCCGACGAAATGTCTGAA
TGCCTACAGTTATTTGGCCATCCTGAAAAGTGCAACTTATCCTGACGTCTCGAGGGACGG
AAAAGTTACCGAAGTCCAAGGAATGAGTCACTTTGCTCAAATTTGATGAGTAATATCAGG
TGTCATGAAACCCAGTTTCGAAGGAGAGGGGAGGGGGCGTCAGATCTGCAGACGGAAGCA
GGCCGCTCCGGATTGGATGGCGAGACCTCGATTTTCCTAAAATTGCGTCATTTAGAACCC
AATTGGGTCCAGATGTTATGGGCATCGACGAGTTACCGTCTCGGAAACTCTCAATCACGC
AAGCGAAAGGAGAGGAGGCGGCTAATTAAATATTGAGCAGAAAGTCGCGTGGGGAGAATG
TCACGTGGGTCTGGAGGCTCAAGGAGGCTGGGATAAATACCGCAAGGCACTGAGCAGGCG
AAAGAGCGCGCTCGGACCTCCTTCCCGGCGGCAGCTACCGAGAGTGCGGAGCGACCAGCG
TGCGCTCGGAGGAACCAGAGAAACTCAGCACCCCGCGGGACTGTCCGTCGCAGTAAGTGC
CCGCGCGGTGCTGGCCGCGGCTGCCCGGGTCACCCCGCCCCGCATCTGTCCGAGGTGGCC
GCGCTGGGGGCGCCGCTGCGGCGAGGGACAGTGGGGAGACTGGCTTCCCAAACGCCAACG
CCCCTCTTTGTCTTCCACCTGCAGAGTTTCCTGGTTTGAAGGTGTGGGTTGGTGGGTTAG
```

- ***How to retrieve a gene transcript sequence***

e.g.11 *Retrieve the TAC1-201 transcript sequence.*

Make sure to check (Transcript: TAC-201 tab) at the top of the page

cDNA sequence can be retrieved from ***the sidebar →Sequence→ cDNA***



Also, protein sequence of a transcript can be retrieved in the same manner

Transcript: TAC-201 tab → ***the sidebar →Sequence→ protein***

- ***SNPs associated with each gene***

e.g.12 *Find two SNPs associated with the gene.*

Make sure to check (Gene tab) at the top of the page

Gene sequence can be found from ***the sidebar → Genetic variation → Variant table***