

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.

o 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

The image shows four screenshots of the Ensembl search interface with numbered annotations:

- 1:** The search dropdown menu is set to "All species".
- 2:** The dropdown menu is open, and "Human" is selected.
- 3:** The search dropdown is set to "Human" and the search term "TAC1" is entered in the search box.
- 4:** The "Restrict category to:" dropdown menu is open, and "Gene" is selected.
- 5:** The search results page for "TAC1 (Human Gene)" is shown, with the gene name highlighted.

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

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq Match	Flags
TAC1-201	ENST00000319273.10	1061	129aa	Protein coding	CCDS5649.g	P20366.g	NM_003182.3.g	TSL:1 GENCODE basic APPRIS P1 MANE Select v0.9
TAC1-202	ENST00000498867.4	1011	114aa	Protein coding	CCDS5651.g	P20366.g	-	TSL:3 GENCODE basic
TAC1-203	ENST00000395048.8	975	111aa	Protein coding	CCDS5650.g	P20366.g	-	TSL:5 GENCODE basic
TAC1-204	ENST00000481437.1	607	No protein	Retained intron	-	-	-	TSL:3
TAC1-205	ENST00000495916.1	522	No protein	Retained intron	-	-	-	TSL:3

Note:

G → gene

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

- *Splice variants* are more than one transcript of a gene due to alternative splicing. Therefore 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?

Gene: TAC1 ENSG00000006128

Description: tachykinin precursor 1 [Source:HGNC Symbol;Acc:HGNC:11517.g]

Gene Synonyms: NKNA, NPK, TAC2

Location: Chromosome 7: 97,732,084-97,740,472 forward strand.
GRCh38.CM000669.2

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

Chromosome 7: 97,732,084-97,740,472

Assembly exceptions: chromosome 7

Region in detail

Scroll: [arrows] Track height: [arrow] Drag/Select: [arrows]

Chromosome bands: 97.40 Mb, 97.60 Mb, 97.80 Mb, 98.00 Mb, 98.20 Mb

Centis: p14.3, p14.1, q21.11, q22.1, q21.1, q31, q32, q33

Genes: AP152P1, RN7SKP104, TAC1, AC005326.1, MIR5682A1, MIR5682A1, AC079781.2, RN7SL478P, <ABNS AC079781.2> RN7SL478P, <AC079781.3> AC079781.4, <RPESAP29> AC004967.1, AC079781.1, <QCM2 AC079781.1> <QCM2 ORF7E7P, SNRPCP9> <AC004967.2 AC079781.3> <ORF7E8P, MIR5692C2

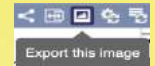
Regulatory Build: [various colored bars]

Gene Legend: merged Ensembl/Havana, RNA gene, pseudogene

Regulation Legend: CTCF, Open Chromatin, Promoter Flank, Enhancer, Promoter, Transcription Factor Binding Site

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	CCDS5649.4
TAC1-202	ENST00000346867.4	1011	114aa	Protein coding	CCDS5651.4
TAC1-203	ENST00000350485.8	975	111aa	Protein coding	CCDS5650.2
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?

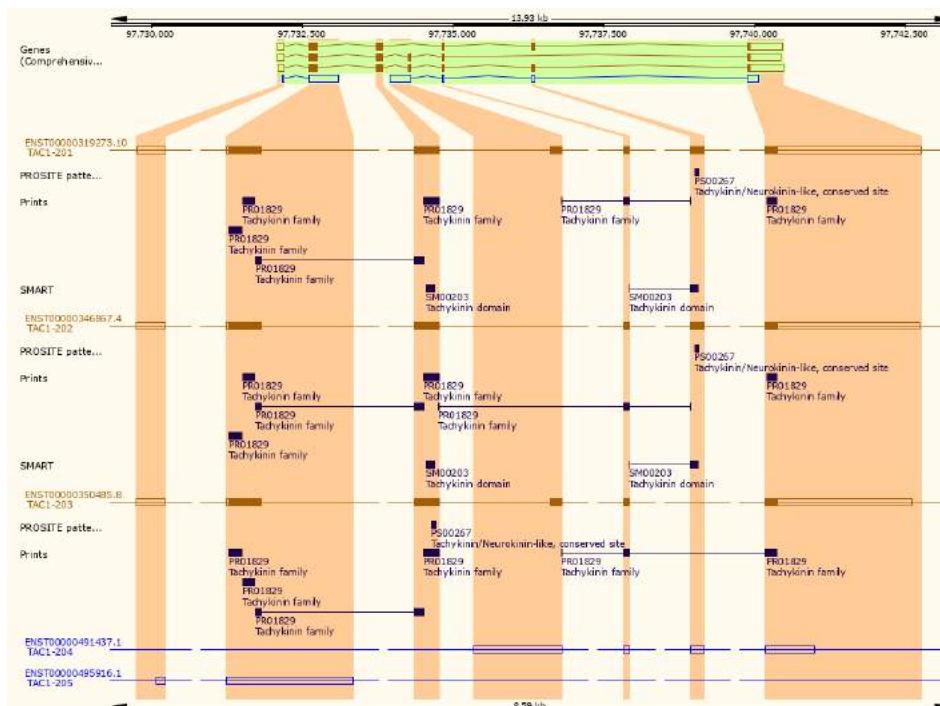
TAC1-201, TAC1-202, and TAC1-203

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).

• **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**

Gene: TAC1 ENSG00000006128

Gene-based displays

- Summary
- Sequence
- Comparative Genomics
- Ontologies
 - GO: Cellular component
 - GO: Molecular function
 - GO: Biological process
- Phenotypes**

Phenotypes

Phenotypes, diseases and traits associated with this gene ENSG00000006128

None found.

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

Phenotype, disease and trait	Source(s)	Number of variants	Show/Hide details
ALL variants with a phenotype annotation	*	21	Show
Adventurousness	NHGRI-EBI GWAS catalog	1	Show
Adverse response to radiation therapy	NHGRI-EBI GWAS catalog	1	Show
Asthma	dbGaP	1	Show
Blood pressure	dbGaP	1	Show
Blood protein levels	NHGRI-EBI GWAS catalog	1	Show
Body Height	dbGaP	2	Show
Body Mass Index	dbGaP	1	Show
Cholesterol, LDL	dbGaP	1	Show
Coronary Artery Disease	dbGaP	3	Show

• **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**

Gene: TAC1 ENSG00000006128

Gene-based displays

- Summary
- Sequence**
- Comparative Genomics
- Ontologies
 - GO: Cellular component
 - GO: Molecular function

Marked-up sequence

Download sequence | BLAST this sequence

Exons **TAC1 exons** All exons in this region

Markup loaded

```
>chromosome:GRCh38:7:97731484:97741072:1
AGGAAAGCCAGTATTTCGCGTTGATTTAGAAGAGGGATGTTCTGGTTATAGAACGATGCT
GTGTCTCAGAAACACTTAAATACTATTAAGCTAGAAATAGAAGGGAAAATAATGCTTCCC
CGCATCTCCCTCAAGTGTAGTCCTCTTTTTTAGCCTGATTTCCGACGAAATGCTGAA
TGCCCTACAGTTAATTTGGCCATCCTGAAAAGTGCAACTTATCCTGACGCTCTGAGGGACGG
AAAAGTTACCGAAGTCCAAGGAATGAGTCACTTTGCTCAAATTTGATGAGTAATATCAGG
TGTATGAAACCCAGTTTCGAAGGAGAGGGGAGGGGCGTCAGATCTGCAGACGGAAGCA
GGCCGCTCCGATTGGATGGCGAGACCTCGATTTTCCTAAAATTCGCTCAATTTAGAACC
AATTGGGTCCAGATGTTATGGGCATCGACGAGTTACCGTCTCGGAAACTCTCAATCACGC
AAGCGAAGGAGAGGAGGCGGCTAATTAATATTGAGCAGAAAATCSCGTGGGAGAATG
TCACGTGGGTCTGGAGGCTCAAGGAGGCTGGGATAAATACCCGAAGGCACTGAGCAGGG
AAAGAGCGCCTCGGACCTCCTTCCCGCGGCAGCTACCGAGAGTGGGAGCGACCAAGG
TGCGCTCGGAGGAACAGAGAACTCAGCACCCCGGGACTGTCCGTCGCAGTAAGTGC
CCGGCGGTGCTGGCCGCGGCTGCCCGGTCACCCGCCCGCATCTGTCCGAGGTGGCC
CGCTGGGGGCGCCGCTGCGGCGAGGACAGTGGGAGACTGGCTTCCCAAACGCCAACG
CCCTCTTTGTCTCCACCTGCAGATTTCTGGTTGAAGTGTGGTTGGTGGTTAG
```

• **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

Gene: TAC1 ENSG00000006128

Gene-based displays

- [-] Summary
- [-] Sequence
- [-] Comparative Genomics
- [-] Ontologies
- [-] Phenotypes
- [-] Genetic Variation
 - Variant table**
 - [-] Variant image
 - [-] Structural variants

Variant table

This table shows known variants for this gene. Use the 'Consequence Type' filter to view a subset of these.

Filter: Global MAF: All SIFT: All PolyPhen: All Consequences: All Filter Other Columns

Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Consequence Type	AA	AA coord	SI FT	PolyPhen	CA DD	RE VEL	MetaLR	Mutation Assessor	Transcript
rs894460002	7:97732088	A/G	-	SNP	dbSNP	-	-	5 prime UTR variant	-	-	-	-	-	-	-	-	ENST00000319273.10
rs112438085	7:97732095	T/A/C/G	0.002 (C)	SNP	dbSNP	AD	-	5 prime UTR variant	-	-	-	-	-	-	-	-	ENST00000319273.10