

## 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 under "Favourite species".
- 3:** 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, including the gene ID ENSG00000006128 and a description of the gene.

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

Gene: **TAC1** ENSG00000006128 ↑ Gene Name

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

Gene Synonyms: ENSEMBL identifier: 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. ↑

Transcripts: Hide transcript table ↑ # of transcripts

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq Match	Flags
TAC1-201	ENST00000319273.10	1061	129aa	Protein coding	CCDS5649	P20366	NM_003182.3	TSL:1 GENCODE basic APPRIS P1 MANE Select v0.9
TAC1-202	ENST00000346867.4	1011	114aa	Protein coding	CCDS5651	P20366	-	TSL:3 GENCODE basic
TAC1-203	ENST00000350485.8	975	111aa	Protein coding	CCDS5650	P20366	-	TSL:5 GENCODE basic
TAC1-204	ENST00000491437.1	607	No protein	Retained intron	-	-	-	TSL:3
TAC1-205	ENST00000495916.1	522	No protein	Retained intron	-	-	-	TSL:3

# Of basepairs    # Of amino acids    Transcript type

**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]

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

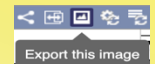
Region in detail

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	<a href="#">ENST00000319273.10</a>	1061	<a href="#">129aa</a>	Protein coding	<a href="#">CCDS5649.1</a>
TAC1-202	<a href="#">ENST00000346867.4</a>	1011	<a href="#">114aa</a>	Protein coding	<a href="#">CCDS5651.1</a>
TAC1-203	<a href="#">ENST00000350485.8</a>	975	<a href="#">111aa</a>	Protein coding	<a href="#">CCDS5650.1</a>
TAC1-204	<a href="#">ENST00000491437.1</a>	607	No protein	Retained intron	-
TAC1-205	<a href="#">ENST00000495916.1</a>	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?

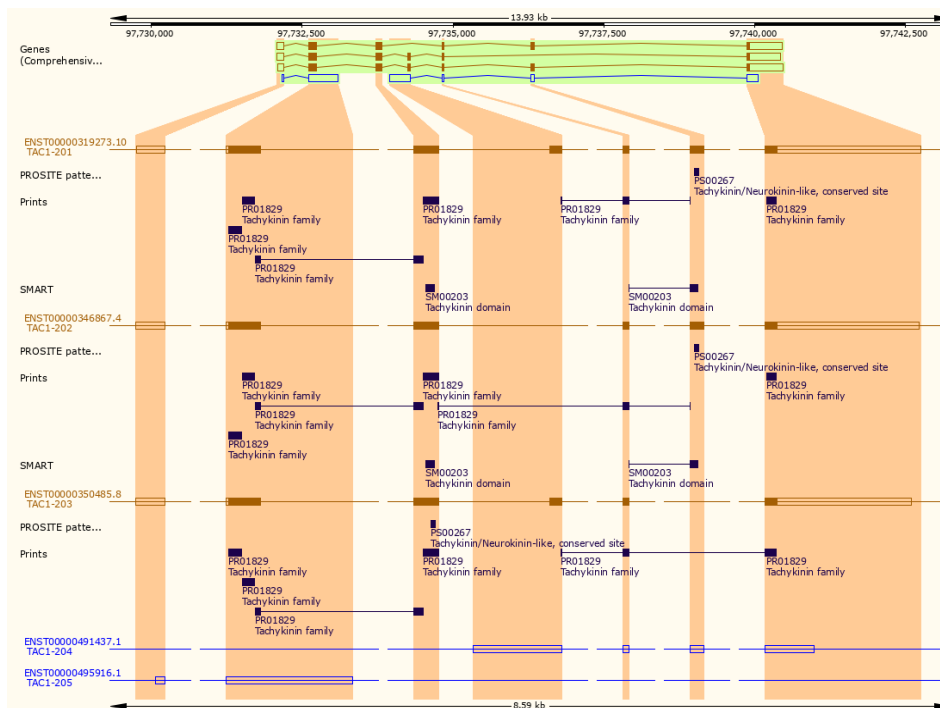
TAC1-201, TAC1-202, and TAC1-203

**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?

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	878	Longest transcript	Protein coding	CCDS5650
TAC1-204	ENST00000491437.1	607	No protein	Retained intron	-
TAC1-205	ENST00000495916.1	522	No protein	Retained intron	-

<b>Transcript: TAC1-201</b>	ENST00000319273.10
<b>Description</b>	tachykinin precursor 1 [Source:HGNC Symbol;Acc:HGNC:11517]
<b>Gene Synonyms</b>	NKNA, NPK, TAC2
<b>Location</b>	Chromosome 7: 97,732,086-97,740,472 forward strand.
<b>About this transcript</b>	This transcript has <b>7 exons</b> , is annotated with <b>17 domains and features</b> , is associated with <b>2009 variant alleles</b> and maps to <b>408 oligo probes</b> .
<b>Gene</b>	This transcript is a product of gene <a href="#">ENSG00000006128.12</a> <a href="#">Hide transcript table</a>

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

**Summary**

Statistics: Exons: 7, Coding exons: 6, Transcript length: 1,061 bps, Translation length: 129 residues

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

Summary: Exons

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
1	5' upstream sequence	97,732,086	97,732,195	-	-	110	.....ggaggtccaaggagctgggataaataacggcaaggcactgagcaggogaa
2	Exon 1-2	97,732,195	97,732,603	-	0	132	AAGGCGGCGGAGAGAGAGGAGCAAAAGGAGCGCAGCAGCGACTGGTGCAC
3	Intron 2-3	97,732,726	97,733,222	0	1	987	gtcagggccccccgagcggcgcgc.....cttttccgttttttgtctcccag
4	Exon 3-4	97,733,820	97,734,247	1	1	45	ATTCCTCAATTAAACAAGGCTCCCTTTAAAGCCTGATG
5	Intron 4-5	97,734,293	97,734,825	1	1	533	gtaaacattcctataaattctcttatt.....cttgcattttatctctctctag
6	Exon 5-6	97,734,950	97,736,298	1	1	24	GAGCCTACAGTCTCCAGAAA
7	Intron 6-7	97,736,353	97,739,873	1	-	3,521	gtaagtcaaaattatttgcatt.....atcacccctaaatgtatitctccag

**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

→ Therefore, 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.g.8** 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**

The screenshot shows the Ensembl interface for gene TAC1 (ENSG00000006128). In the 'Gene-based displays' sidebar, 'Phenotypes' is highlighted with a red box. Below, the 'Phenotypes' section shows a table of annotations for variants in this gene. A red box highlights the first few rows of the table, with a red '2' indicating the number of entries shown.

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**

The screenshot shows the 'Gene-based displays' sidebar for TAC1. 'Sequence' is highlighted with a red box.

**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
GTGTCTCAGAAACACTTAAATACTATTAAGCTAGAAATAGAAGGGAAAAATAATGCTTCCC
CGCATCTCCCTCAAGTGTAGTCCTCTTTTTTAGCCTGATTTCCGACGAAATGCTGAA
TGCTTACAGTTATTTGGCCATCTGAAAAGTGCAACTTATCCTGACGTCGAGGGACGG
AAAAAGTTACCGAAGTCCAAGGAATGAGTCACTTTGCTCAAATTTGATGAGTAATATCAGG
TGTATGAAACCCAGTTTCGAAGGAGAGGGGAGGGGCGTCAGATCTGCAGACGGAAGCA
GGCCGCTCCGATTGGATGGCGAGACCTCGATTTTCCTAAAATTCGCTCATTTAGAACC
AATTGGGTCCAGATGTTATGGGCATCGACGAGTTACCGTCTCGGAAACTCTCAATCAGC
AAGCGAAGGAGAGGAGGCGGCTAATTAATATTGAGCAGAAAGTCGCGTGGGAGAAATG
TCACGTTGGTCTGGAGGCTCAAGGAGGCTGGGATAAATACCGCAAGGCACTGAGCAGGCG
AAAGAGCGCCTCGACCTCTTCCCGGGCGGACGTACCGAGAGTCCGGAGCCACCGG
TCCGCTCGGAGGAACAGAGAACTCAGCACCCCGGGGACTGTCCTCGCAGTAAGTGC
CCGCGCGGTGCTGGCCGCGGCTGCCCGGGTCAACCCGCCCGCACTCTCCAGGTGGCC
GGCTGGGGGCGCCGCTCGCGCAGGGACAGTGGGGAGACTGGCTTCCCAAACGCCAACG
CCCTCTTTGCTTCCACTGCAGATTTCTGTTTGAAGTGTGGTTGGTGGTTAG
```

• **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	Glo-bal MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA coord	SI FT	Pol y-Phe n	CA DD	RE VEL	Met aLR	Mut atio n Ass ess or	Transcript
<b>rs894460002</b>	7:97732088	A/G	-	SNP	dbSNP	-	-	5 prime UTR variant	-	-	-	-	-	-	-	-	ENST00000319273.10
<b>rs112438085</b>	7:97732095	T/A/C/G	0.002 (C)	SNP	dbSNP	AD	-	5 prime UTR variant	-	-	-	-	-	-	-	-	ENST00000319273.10