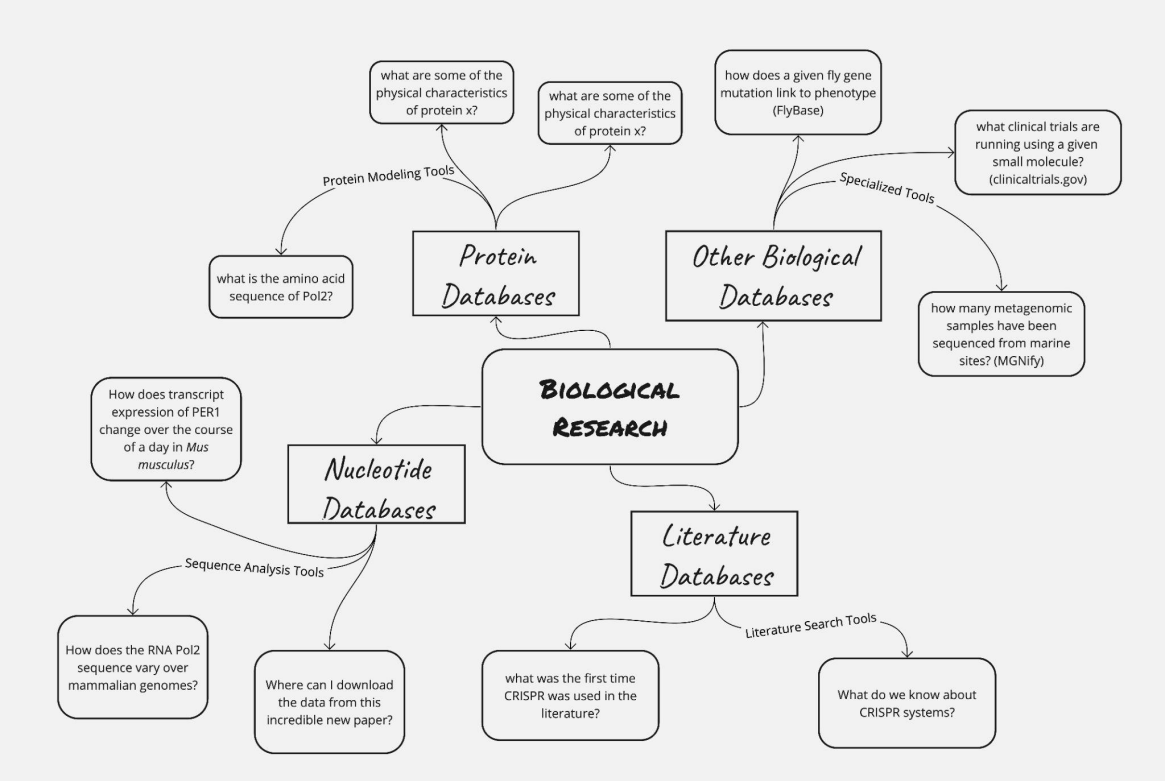
Online Databases

Adrian Filips & Phoebe Oldach

It's a jungle out there !
Categories - Publications and Life/Organism based databases



- Different jurisdictions / silos
- Different data, structured in different ways
- Different tools

Specialized Databases: These are specific to certain organisms, diseases, or types of data. Examples include:

- **FlyBase**: A database for Drosophila genetics and molecular biology.
- **The Cancer Genome Atlas:** Provides genomic and molecular data on various types of cancer.
- **Saccharomyces Genome Database**: a database for yeast genomes and mol bio

Taxonomic and Species Databases: These databases provide information on species classification and taxonomy. Examples include:

- **ITIS** (Integrated Taxonomic Information System): Taxonomic information on plants, animals, fungi, and microbes.
- **The Plant List**: A comprehensive botanical database

Ecological and Environmental Databases: Contain data on ecosystems, biodiversity, and environmental conditions. Examples include:

- **GBIF** (Global Biodiversity Information Facility): Provides access to data about all types of life on Earth.
- **DataONE**: A repository of environmental and ecological data.

What is a database and what content do we expect for bio databases?

Columns:

Name

Source

Organism

Accession – similar to version External references

Treasure column containing the sequence

Central Dogma is DNA→RNA→Proteins so we expect to have DNA (AKA nucleotide/genomic) and RNA and protein databases.

RNA information is indirectly stored in DNA databases, therefore we only focus on **Protein databases and DNA databases** in this course - they are the ones we will need them for our homework

Proteins are chains of Amino-acids, that can be concatenated in long sequences and they have a “Beginning”-called N Terminal and an “End” called C Terminal and elongation occurs only at the end. Each AA uses a letter, for instance Methionine is M. So to capture information on a protein we just need a simple **oriented** sequence of letters like **MALWMRLL**. growing to the →

That **sequence c**olumn and stores the amino-acid names in order starting with N terminal. That is what we will search and download as a text file to further process it in other tools to do our magic.

Nucleotide Databases: These databases contain sequence data for DNA and RNA. Examples include:

- **GenBank**: A genetic sequence database providing an annotated collection of all publicly available DNA sequences.
- **EMBL Nucleotide Sequence Database**: Managed by the European Molecular Biology Laboratory, it collects nucleotide sequences.

Metabolic and Pathway Databases: These focus on metabolic pathways and interactions within a cell. Examples include:

- **KEGG (Kyoto Encyclopedia of Genes and Genomes)**: A database resource for understanding high-level functions of the biological system.
- **Reactome**: Provides curated pathways of many biological processes.

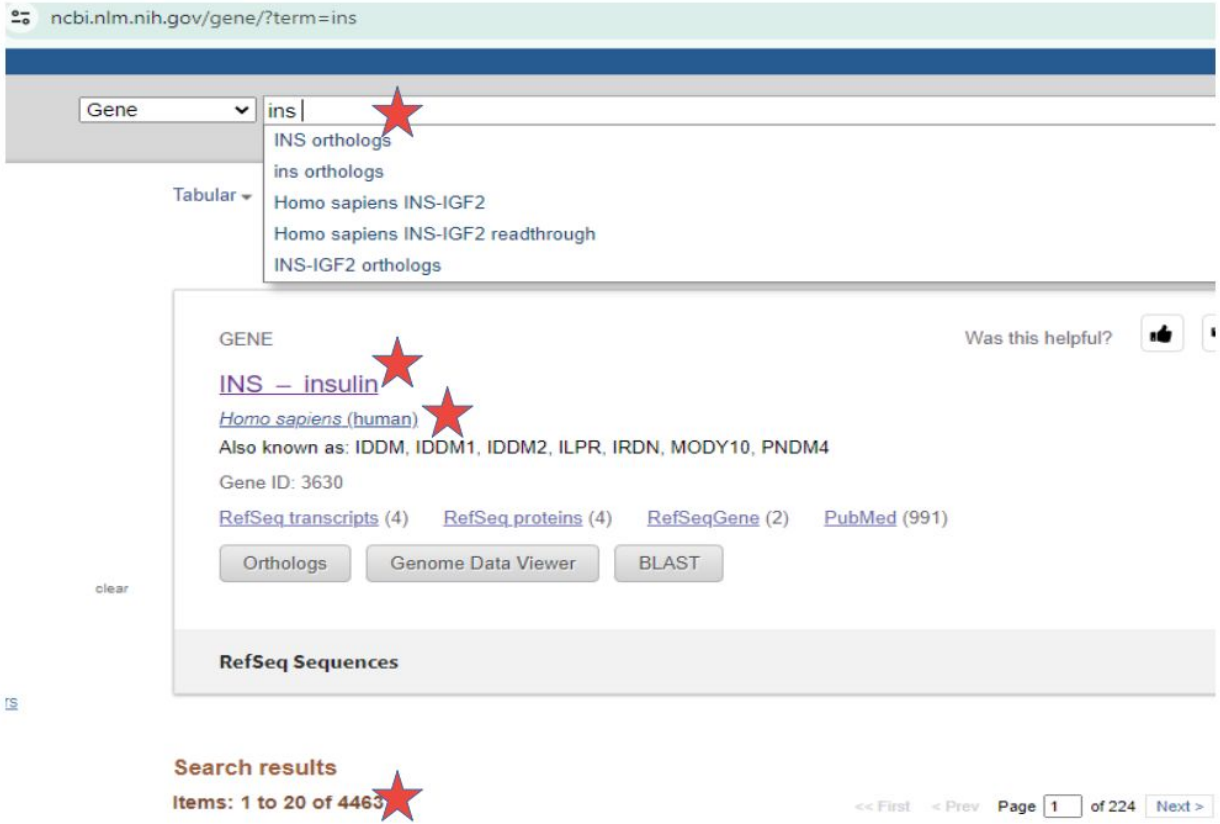
Protein Databases: Focused on proteins and store sequences, structures, and functions:

- **UniProt**: Protein sequence and annotation data.
- **Protein Data Bank (PDB)**: A database for the 3D structural data of large biological molecules.
- **SwissProt**: Protein sequence and functional information

DNA is made of 4 molecules A,T,G,C and like amino-acids they have a beginning and an end called **5' and 3'** and can be linked in long **sequences** called strands
In order to resist breakage the DNA strands are doubled by using **complementary bases oriented in reverse direction**. Only one **strand sequence** is stored in the DNA database and that is called coding strand and it's not !!! the one read by RNA polymerase!!

DNA Databases sometimes called Nucleotide/Genomic **:**

- **NCBI Genome**: National Center for Biotechnology Info complete genome sequences
- **Ensembl/BLAST**: A genomic browser and tools
- **EMBL**: Collection of nucleotide sequences by European Bioinformatics Institute


ncbi.nlm.nih.gov/gene/?term=ins
Gene
ins
INS orthologs
ins orthologs
Homo sapiens INS-IGF2
Homo sapiens INS-IGF2 readthrough
INS-IGF2 orthologs
Tabular
GENE
Was this helpful?
INS – insulin
Homo sapiens (human)
Also known as: IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10, PNDM4
Gene ID: 3630
RefSeq transcripts (4)
RefSeq proteins (4)
RefSeqGene (2)
PubMed (991)
Orthologs
Genome Data Viewer
BLAST
clear
RefSeq Sequences
Search results
Items: 1 to 20 of 4463
<< First
< Prev
Page 1 of 224
Next >

sulin [*Homo sapiens* (human)]

Download Datasets

8630, updated on 16-Jan-2024

mary

Official Symbol	INS provided by HGNC
fficial Full Name	insulin provided by HGNC
Primary source	HGNC:HGNC:6081
See related	Ensembl:ENSG00000254647 MIM:176730; AllianceGenome:HGNC:6081
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	IDDM; ILPR; IRDN; IDDM1; IDDM2; PNDM4; MODY10
Summary	This gene encodes insulin, a peptide hormone that plays a vital role in the regulation of carbohydrate and lipid metabolism. After removal of the precursor signal peptide, proinsulin is post-translationally cleaved into three peptides: the B chain and A chain peptides, which are covalently linked via two disulfide bonds to form insulin, and C-peptide. Binding of insulin to the insulin receptor (INSR) stimulates glucose uptake. A multitude of mutant alleles with phenotypic effects have been identified, including insulin-dependent diabetes mellitus, permanent neonatal diabetes diabetes mellitus, maturity-onset diabetes of the young type 10 and hyperproinsulinemia. There is a read-through gene, INS-IGF2, which overlaps with this gene at the 5' region and with the IGF2 gene at the 3' region. [provided by RefSeq, May 2020]
Expression	Restricted expression toward pancreas (RPKM 671.7) See more
Orthologs	mouse all

NEW Try the new Gene table

Try the new Transcript table

omic context

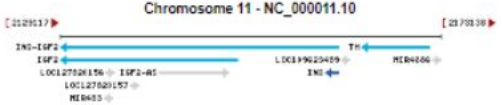
on: 11p15.5

See INS in Genome Data Viewer

ount: 3

ation release	Status	Assembly	Chr	Location
023_10	current	GRCh38.p14 (GCF_000001405.40)	11	NC_000011.10 (2159779..2161209, complement)
023_10	current	T2T-CHM13v2.0 (GCF_009914755.1)	11	NC_060935.1 (2247427..2248857, complement)
0220307	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (2181009..2182439, complement)

Chromosome 11 - NC_000011.10

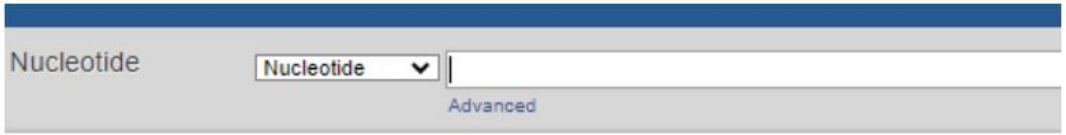


omic regions, transcripts, and products

Go to reference sequence details

ic Sequence: NC_000011.10 Chromosome 11 Reference GRCh38.p14 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank



FASTA

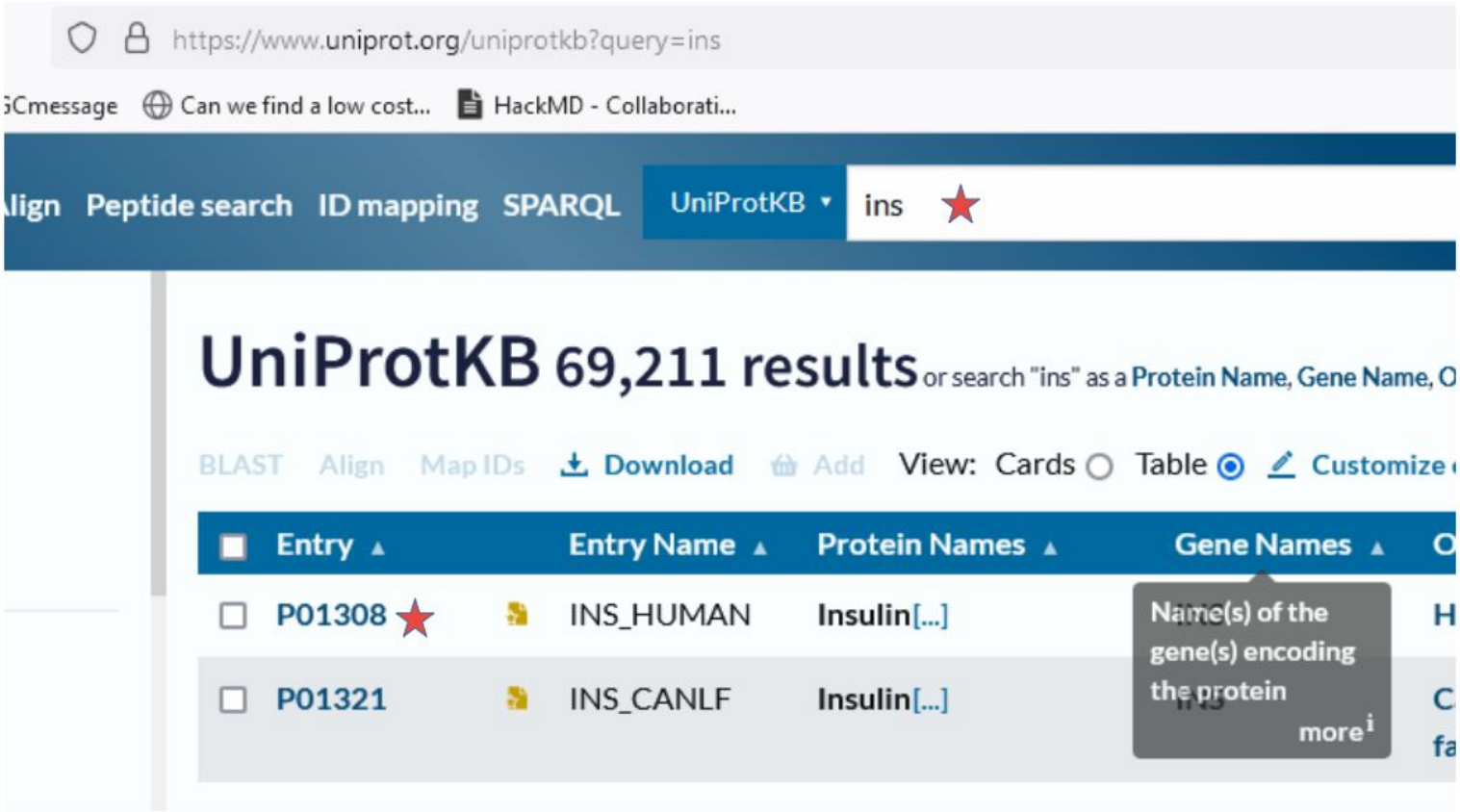
Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly

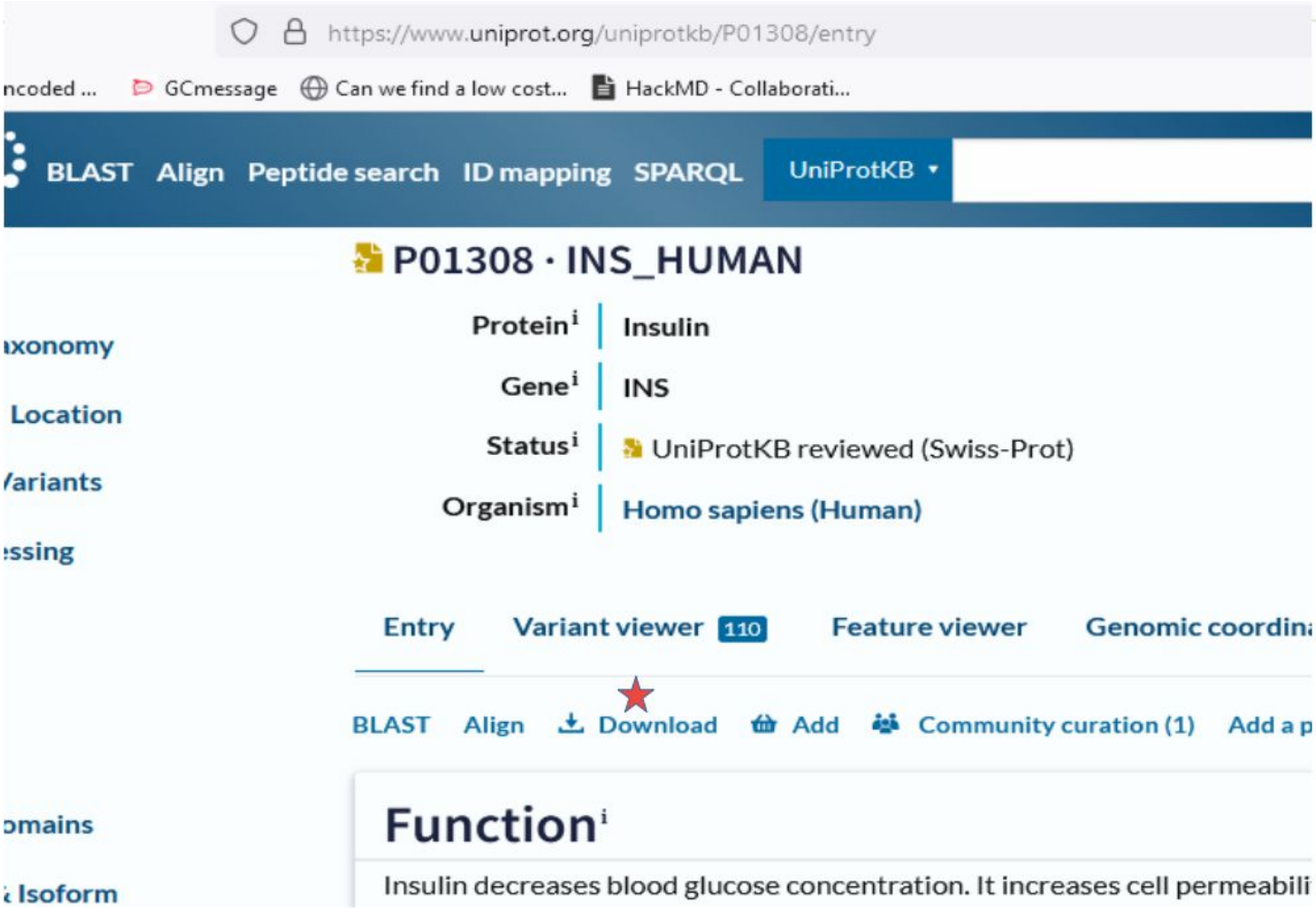
NCBI Reference Sequence: NC_000011.10

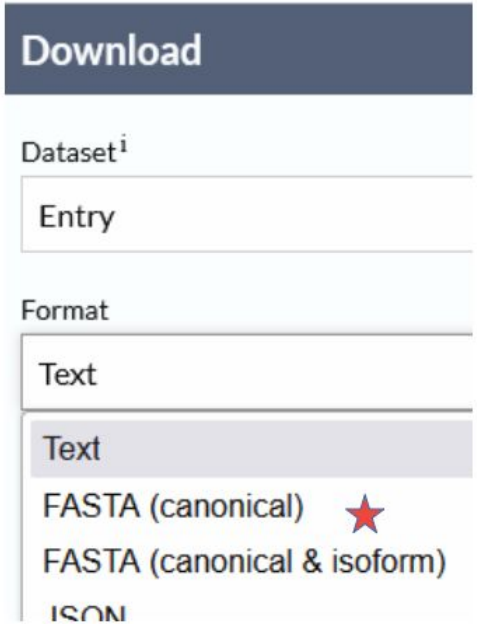
GenBank Graphics

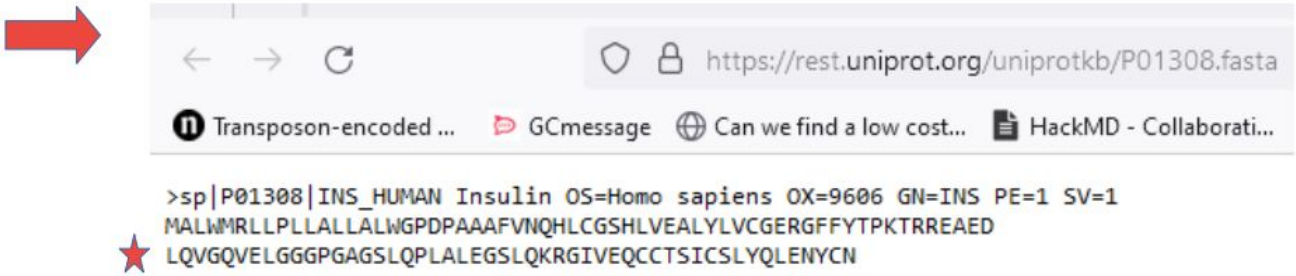
>NC_000011.10:c2161209-2159779 Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGT
GGGCTCAGGATTCCAGGGTGGCTGGACCCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCG
TGAAGCATGTGGGGGTGAGCCCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGC
CTGTCTCCCAGATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTG
GCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAG
CTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCT
GCAGGGTGAGCCAACTGCCCATTGCTGCCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCAC
CCAGCATGGGCAGAAGGGGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAG
TTCTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGAGGACGGTG
TTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCCACTCGCCCCTCAAACAAA
TGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAGATTCAAGTGTTTTGTTAAGTAAAGTCCT
GGGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGGGCGAACACCCCATCACGCCCGGAGGA
GGGCGTGGCTGCCTGCCTGAGTGGGCCAGACCCCTGTCGCCAGGCCTCACGGCAGCTCCATAGTCAGGAG
ATGGGGAAGATGCTGGGGACAGGCCCTGGGGAGAAGTACTGGGATCACCTGTTCAGGCTCCCACTGTGAC
GCTGCCCCGGGGCGGGGGAAGGAGGTGGGACATGTGGGCGTTGGGGCCTGTAGGTCCACACCCAGTGTGG
GTGACCCTCCCTCTAACCTGGGTCCAGCCCGGCTGGAGATGGGTGGGAGTGCGACCTAGGGCTGGCGGGC
AGGCGGGCACTGTGTCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCGCCGCTGTTCCGGAACCTGC
TCTGCGCGGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGC
CCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCT
CTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGCCTCCTGCACC
GAGAGAGATGGAATAAAGCCCTTGAACCAGC

```
                     ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL
                     YQLENYCN"
     CDS             join(239..425,1213..1358)
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10;
                     PNDM4"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /codon_start=1
                     /product="insulin preproprotein"
                     /protein_id="NP_000198.1"
                     /db_xref="CCDS:CCDS7729.1"
                     /db_xref="Ensembl:ENSP00000370731.5"
                     /db_xref="GeneID:3630"
                     /db_xref="HGNC:HGNC:6081"
                     /db_xref="MIM:176730"
                     /translation="MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG
                     ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL
                     YQLENYCN"
     CDS             join(239..425,1213..1358)
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10;
                     PNDM4"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /codon_start=1
                     /product="insulin preproprotein"
                     /protein_id="NP_001278826.1"
                     /db_xref="CCDS:CCDS7729.1"
                     /db_xref="GeneID:3630"
                     /db_xref="HGNC:HGNC:6081"
                     /db_xref="MIM:176730"
                     /translation="MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG
                     ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL
                     YQLENYCN"
ORIGIN
        1 agccctccag gacaggctgc atcagaagag gccatcaagc aggtctgttc caagggcctt
       61 tgcgtcaggt gggctcagga ttccagggtg gctggacccc aggccccagc tctgcagcag
      121 ggaggacgtg gctgggctcg tgaagcatgt gggggtgagc ccaggggccc caaggcaggg
      181 cacctggcct tcagcctgcc tcagccctgc ctgtctccca gatcactgtc cttctgccat
```


https://www.uniprot.org/uniprotkb?query=ins
Can we find a low cost...
HackMD - Collaborati...
lign
Peptide search
ID mapping
SPARQL
UniProtKB
ins
UniProtKB 69,211 results
or search "ins" as a Protein Name, Gene Name, O
BLAST
Align
Map IDs
Download
Add
View: Cards
Table
Customize
Entry
Entry Name
Protein Names
Gene Names
P01308
INS_HUMAN
Insulin[...]
P01321
INS_CANLF
Insulin[...]
Name(s) of the gene(s) encoding the protein
more


https://www.uniprot.org/uniprotkb/P01308/entry
ncoded ...
GCmessage
Can we find a low cost...
HackMD - Collaborati...
BLAST
Align
Peptide search
ID mapping
SPARQL
UniProtKB
P01308 · INS_HUMAN
Protein
Insulin
Gene
INS
Status
UniProtKB reviewed (Swiss-Prot)
Organism
Homo sapiens (Human)
axonomy
Location
Variants
essing
Entry
Variant viewer 110
Feature viewer
Genomic coordina
BLAST
Align
Download
Add
Community curation (1)
Add a p
omains
Isoform
Function
Insulin decreases blood glucose concentration. It increases cell permeabili


Download
Dataset[i]
Entry
Format
Text
Text
FASTA (canonical)
FASTA (canonical & isoform)
JSON


https://rest.uniprot.org/uniprotkb/P01308.fasta
Transposon-encoded ...
GCmessage
Can we find a low cost...
HackMD - Collaborati...
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

https://en.wikipedia.org/wiki/List_of_biological_databases