Practical 1 -corrections

Protein sequence databases and sequence annotation

Swiss Institute of Bioinformatics

http://education.expasy.org/cours/SIB_UniProtKB_2020/

From nucleic acid sequence databases to protein sequence databases

Look at this entry (https://www.ncbi.nlm.nih.gov/nuccore/X02158)

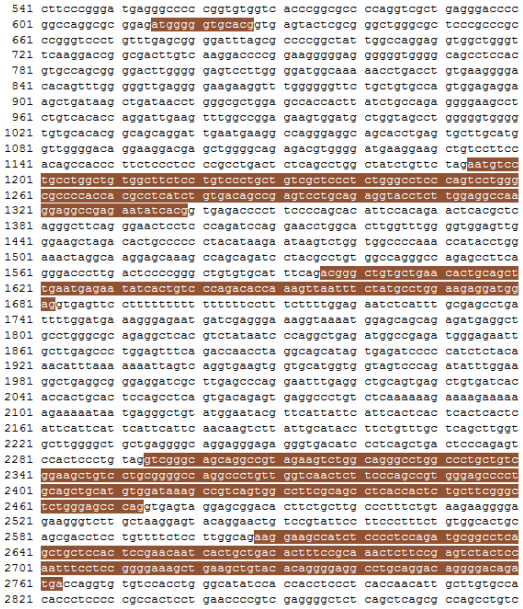
- Which server? Which database? Which accession number?

  NCBI, GenBank, X02158

- What is the main type of data?

  Human gene for erythropoietin

- Look at the CoDing Sequence (CDS): click on 'CDS': how many exons? (5 exons)
- Can you find any information on the accuracy of the CDS? No, you have to go back to the publication PMID: **3838366**



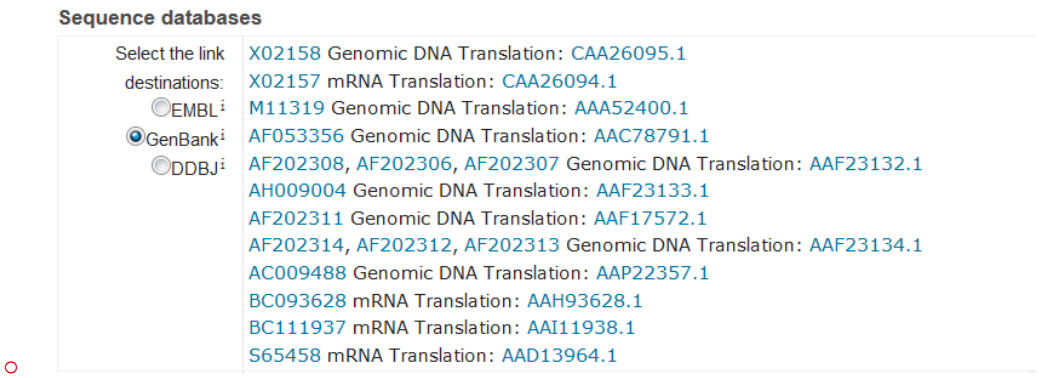
- Follow the link to UniProtKB

  https://www.uniprot.org/uniprot/P01588

- From UniProtKB: go back to the GenBank entry (*Cross-reference section*)

NB: notice that there are many cross-references to GenBank entries which have been used to 'construct' the UniProtKB/Swiss-Prot sequences (different isoforms)

  - Can you find a link to the entries X02158 ?



Look at this entry (https://www.ncbi.nlm.nih.gov/nuccore/NM_000799)

- Which server? Which database? Which accession number?

  NCBI, RefSeq, NM_000799

- What is the main type of data?

  Homo sapiens erythropoietin (EPO), mRNA

- Look at the ##Evidence-Data-START##: How is this entry related to X02157

  Nucleotide Sequences from 1-597 and 601-1144 have been used to construct the mRNA NM_000799

PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-332	AC009488.5	61278-61609	
	333-929	X02157.1	1-597	
	930-932	S65458.1	248-250	
	933-1476	X02157.1	601-1144	
	1477-1477	AC009488.5	64325-64325	
	1478-1662	X02157.1	1146-1330	

- Follow the link to the protein sequence entry (NP_); from there, follow the link to the corresponding UniProtKB entry (column on the right: 'Related information')

  https://www.ncbi.nlm.nih.gov/protein/62240997

- From the UniProtKB entry (NCBI view), go to the original view of this entry on the UniProt website.

  https://www.ncbi.nlm.nih.gov/protein/119526

  ➔ https://www.uniprot.org/uniprot/P01588

Discover the content of a UniProtKB entry, the differences between Swiss-Prot and TrEMBL and the source(s) of annotation

Look at this UniProtKB entry (P04150)

- 'Header section'

What is the status of this entry: reviewed by a biocurator, or unreviewed?

What is the evidence for the existence of the protein? (experimental evidence at protein level)

Have a look at the types of evidence that supports the existence of a protein.

- 'Names &Taxonomy' section

What are the name(s) of the gene and the name(s) of the protein?

Protein names	*Recommended name:* **Glucocorticoid receptor** ▪ *Short name:* GR *Alternative name(s):* • Nuclear receptor subfamily 3 group C member 1
Gene names	*Name:* NR3C1 Synonyms: GRL

What is the 'Taxonomic identifier' (TaxId)? **9606**

Does this entry belong to a proteome, and if so, what is the identifier of that proteome?
UP000005640

How many proteins in this proteome 74788 ? How many proteins are encoded by chromosome 5?
3100

Organism	Homo sapiens (Human)
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Homo
Proteomes	UP000005640 Component: Chromosome 5

https://www.uniprot.org/proteomes/UP000005640

Proteomes - Homo sapiens (Human)

None
Overview
Components
Publications

Map to
UniProtKB (74,788)
Reviewed (20,352) Swiss-Prot
Unreviewed (54,436) TrEMBL

Overview

Status	Reference proteome
Proteins	74,788
Gene count	20,605 - Download one protein sequence per gene (FASTA)
Proteome ID	UP000005640
Taxonomy	9606 - Homo sapiens
Last modified	November 5, 2019
Genome assembly and annotation	GCA_000001405.27 from Ensembl full
Busco	C:99.9%[S:0.9%,D:99%],F:0.1%,M:0%,n:6192
Completeness	Close to Standard

Homo sapiens (Homo sapiens sapiens) or modern humans are the only living species of the evolutionary branch of great apes known as hominids. Divergence of early humans from chimpanzees and gorillas is estimated to have occurred between 4 and 8 million years ago. The genus Homo (Homo habilis) appeared in Africa around 2.3 million years ago and shows the first signs of stone tool usage. The exact lineage of Homo species ie: H. habilis/H. ergaster to H. erectus to H. rhodesiensis/H.heidelbergensis to H. sapiens is still hotly disputed. However, continuing evolution and in particular larger brain size and complexity culminates in Homo sapiens. The first anatomically modern humans appear in the fossil record around 200,000 years ago. Modern humans migrated across the globe essentially as hunter-gatherers until around 12,000 years ago when the practice of agriculture and animal domestication enabled large populations to grow leading to the development of civilizations. Overall life expectancy in Europe is 81 years.

Components

Download | View all proteins

Component name	Genome Accession(s)	Component representation	Proteins
Chromosome 1	CM000663		6028
Chromosome 2	CM000664		4805
Chromosome 3	CM000665		4376
Chromosome 4	CM000666		2797
Chromosome 5	CM000667		3100

- 'Sequences' section

How many different protein sequences (isoforms) are available for this gene?

Sequences (16+)

Sequence status: Complete.

This entry describes **16** isoforms produced by **alternative splicing** and **alternative initiation**. Align | Add to basket

This entry has 16 described isoforms and 3 potential isoforms that are computationally mapped. Show all | Align All

How many in UniProtKB/Swiss-Prot? How many in UniProtKB/TrEMBL (computationaly mapped)?
Look at the length of the different protein sequences

Computationally mapped potential isoform sequences

There are 3 potential isoforms mapped to this entry. BLAST | Align | Show all | Add to basket

Entry	Entry name	Protein names	Gene names	Length	Annotation
D6RDA9	D6RDA9_HUMAN	Glucocorticoid receptor	**NR3C1**	144	Annotation score:
Q3MSN4	Q3MSN4_HUMAN	Glucocorticoid receptor	**NR3C1**	145	Annotation score:
A0A494C0P1	A0A494C0P1_HUMAN	Glucocorticoid receptor	**NR3C1**	746	Annotation score:

- '**Cross references**' **section** (sequence databases)

Look at the RefSeq cross references: How many entries?

RefSeq

NP_000167.1, NM_000176.2 [P04150-1]
NP_001018084.1, NM_001018074.1 [P04150-1]
NP_001018085.1, NM_001018075.1 [P04150-1]
NP_001018086.1, NM_001018076.1 [P04150-1]
NP_001018087.1, NM_001018077.1 [P04150-1]
NP_001018661.1, NM_001020825.1 [P04150-2]
NP_001019265.1, NM_001024094.1 [P04150-3]
NP_001191187.1, NM_001204258.1 [P04150-8]
NP_001191188.1, NM_001204259.1 [P04150-11]
NP_001191189.1, NM_001204260.1 [P04150-12]
NP_001191190.1, NM_001204261.1 [P04150-13]
NP_001191191.1, NM_001204262.1 [P04150-14]
NP_001191192.1, NM_001204263.1 [P04150-15]
NP_001191193.1, NM_001204264.1 [P04150-16]
XP_005268476.1, XM_005268419.3
XP_005268477.1, XM_005268420.4
XP_005268479.1, XM_005268422.3 [P04150-3]
XP_005268480.1, XM_005268423.3 [P04150-3]
XP_016864886.1, XM_017009397.1
XP_016864887.1, XM_017009398.1

Why are not all isoforms cross-referenced to RefSeq?

RefSeq: one entry for each mRNAs -> several entries can be linked to the the same Swiss-Prot sequence.
RefSeq might contain additional isoforms which are not in UniProtKB/Swiss-Prot and vice-versa (different gene prediction pipelines (RefSeq vs Ensembl prediction pipelines)

- **'Function' section**

What is the function of the protein?

Function

Receptor for glucocorticoids (GC) (PubMed:27120390). Has a dual mode of action: as a transcription factor that binds to glucocorticoid response elements (GRE), both for nuclear and mitochondrial DNA, and as a modulator of other transcription factors. Affects inflammatory responses, cellular proliferation and differentiation in target tissues. Involved in chromatin remodeling (PubMed:9590696). Plays a role in rapid mRNA degradation by binding to the 5' UTR of target mRNAs and interacting with PNRC2 in a ligand-dependent manner which recruits the RNA helicase UPF1 and the mRNA-decapping enzyme DCP1A, leading to RNA decay (PubMed:25775514). Could act as a coactivator for STAT5-dependent transcription upon growth hormone (GH) stimulation and could reveal an essential role of hepatic GR in the control of body growth (By similarity). By similarity | 3 Publications

Where does the information come from?

By similarity (mouse ortholog) and from publications

Compare the **keywords** and the **Gene Ontology** terms relative to the protein's function

Keywords give a 'summary' of the biological knowledge (manually annotated by the UniProt biocurators).
GO give an exhaustive overview of what is known about your gene (ontology) (annotated by different groups, including the model organism databases).

GO - Molecular function

- core promoter binding — Source: CAFA
- DNA-binding transcription activator activity, RNA polymerase II-specific — Source: UniProtKB
- DNA-binding transcription factor activity — Source: UniProtKB
- DNA-binding transcription factor activity, RNA polymerase II-specific — Source: NTNU_SB
- glucocorticoid receptor activity — Source: ProtInc
- Hsp90 protein binding — Source: UniProtKB
- nuclear receptor activity — Source: UniProtKB
- protein kinase binding — Source: ARUK-UCL
- RNA binding — Source: UniProtKB-KW
- RNA polymerase II cis-regulatory region sequence-specific DNA binding — Source: NTNU_SB
- steroid binding — Source: UniProtKB
- steroid hormone binding — Source: UniProtKB
- SUMO binding — Source: CAFA
- zinc ion binding — Source: InterPro

Complete GO annotation on QuickGO ...

GO - Biological process

- apoptotic process — Source: UniProtKB-KW
- cell cycle — Source: UniProtKB-KW
- cell division — Source: UniProtKB-KW
- cellular response to dexamethasone stimulus — Source: CAFA
- cellular response to glucocorticoid stimulus — Source: UniProtKB
- cellular response to steroid hormone stimulus — Source: UniProtKB
- cellular response to transforming growth factor beta stimulus — Source: CAFA
- chromatin organization — Source: UniProtKB-KW
- chromosome segregation — Source: UniProtKB-KW
- negative regulation of transcription by RNA polymerase II — Source: CAFA
- positive regulation of pri-miRNA transcription by RNA polymerase II — Source: ARUK-UCL
- positive regulation of transcription by RNA polymerase II — Source: UniProtKB
- regulation of transcription, DNA-templated — Source: UniProtKB
- signal transduction — Source: ProtInc
- transcription, DNA-templated — Source: CAFA
- transcription by RNA polymerase II — Source: ProtInc
- transcription initiation from RNA polymerase II promoter — Source: Reactome

Complete GO annotation on QuickGO ...

Keywords

Molecular function	Chromatin regulator, DNA-binding, Receptor, RNA-binding
Biological process	Apoptosis, Cell cycle, Cell division, Chromosome partition, Mitosis, Transcription, Transcription regulation

Note that the 3 GO ontologies Molecular Function, Biological Process and Cellular Component, as well as UniProt Keywords are dispatched to the relevant sections of the entry, not all GO terms can be found under "Function").

What is the source of the Gene Ontology (see the blue and yellow tags) and keyword annotations (Swiss-Prot biocurators)?

FAQ: What are the differences between UniProtKB keywords and the GO terms?

- 'PTM/Processing' section

How many phosphorylated sites?

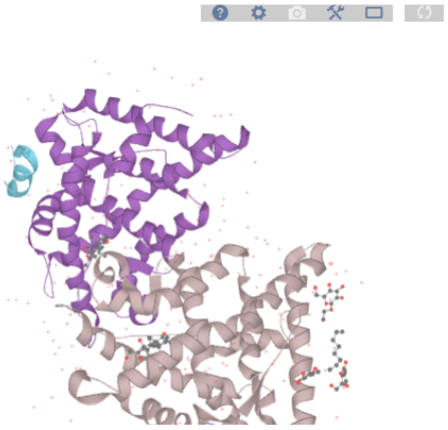
How many sites have been **experimentally proven** to be phosphorylated? the sites associated with a publication are 'experimentally proven.

Amino acid modifications

Feature key	Position(s)	Description
Modified residue	8	Phosphothreonine — Combined sources
Modified residue	23	Omega-N-methylarginine — By similarity
Modified residue	45	Phosphoserine — Combined sources
Modified residue	113	Phosphoserine — By similarity
Modified residue	134	Phosphoserine — Combined sources
Modified residue	141	Phosphoserine — By similarity
Modified residue	203	Phosphoserine — Combined sources — 3 Publications
Modified residue	211	Phosphoserine — 3 Publications
Modified residue	226	Phosphoserine — Combined sources — 1 Publication
Cross-link	258	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2) — Combined sources
Modified residue	267	Phosphoserine — Combined sources
Cross-link	277	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO); alternate — 1 Publication
Cross-link	277	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2); alternate — Combined sources
Cross-link	293	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO); alternate — 1 Publication
Cross-link	293	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2); alternate — Combined sources
Modified residue	404	Phosphoserine; by GSK3-beta — 1 Publication

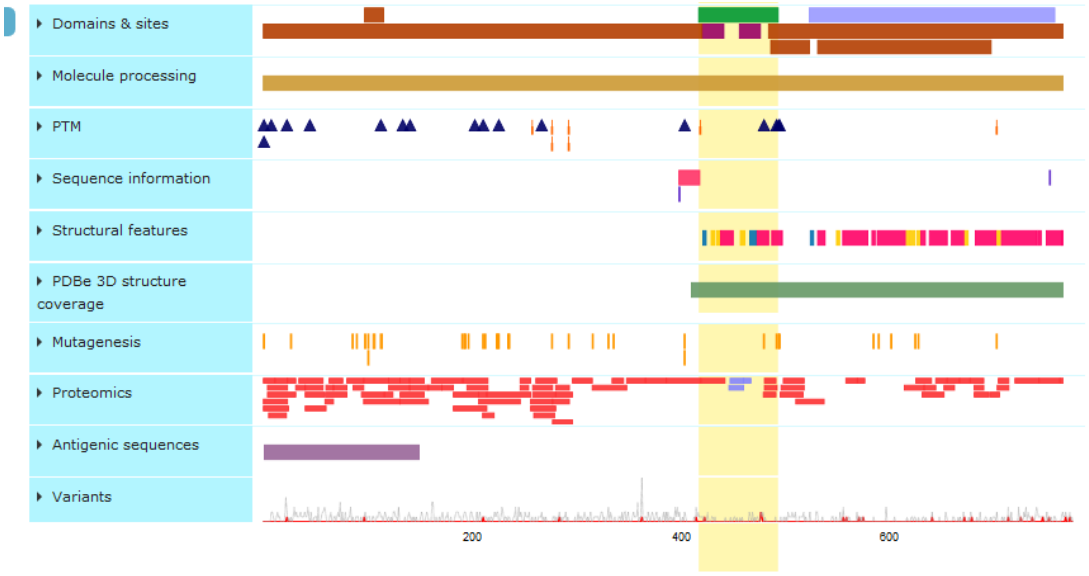
- **'Structure' section** (3D structure databases):
Are there 3D structures are available for this protein? Do they 'cover' the complete protein sequence?
There are several 3D structures for this protein: none of these 3D structures cover the entire protein sequence!

Structure



PDB Entry	Method	Resolution	Chain	Positions	Links
1M2Z	X-ray	2.50 Å	A/D	521-777	PDBe RCSB ... PDBj PDBsum
1NHZ	X-ray	2.30 Å	A	500-777	PDBe RCSB ... PDBj PDBsum
1P93	X-ray	2.70 Å	A/B/C/D	500-777	PDBe RCSB ... PDBj PDBsum
3BQD	X-ray	2.50 Å	A	525-777	PDBe RCSB ... PDBj

- Look at the different tracks of the **Feature viewer** (including 'proteomics' and 'variants')
Look at the DNA binding domain in the 3D structure *(Hint: click on the domain 418-493)*
To which PDB entry does it correspond?
The PDB entry which covers the domain is 4HN5 (417-506)
https://www.uniprot.org/uniprot/P04150/protvista



PDB Entry	Method	Resolution	Chain	Positions	Links
					PDBsu...
3K23	X-ray	3.00 Å	A/B/C	521-777	PDBe RCSB ... PDBj PDBsu...
4CSJ	X-ray	2.30 Å	A	500-777	PDBe RCSB ... PDBj PDBsu...
4HN5	X-ray	1.90 Å	A/B	417-506	PDBe RCSB ... PDBj PDBsu...
4HN6	X-ray	2.55 Å	A/B	417-506	PDBe RCSB ... PDBj PDBsu...

- Look at the same UniProt entry in the 'txt' format (via the “Format” button).

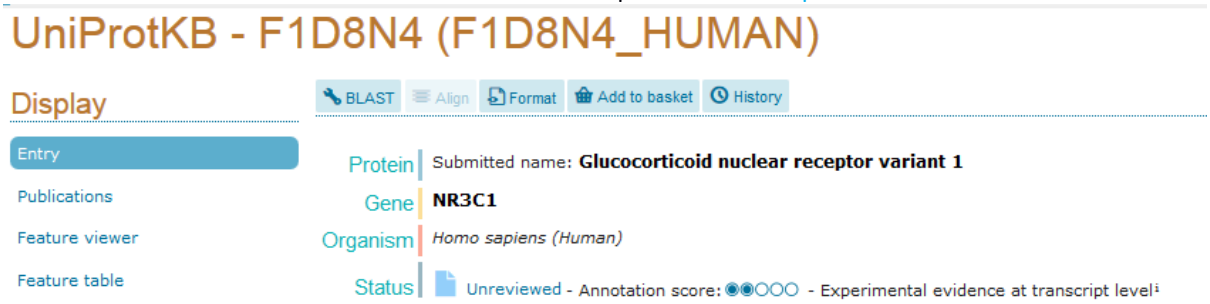
```
ID   GCR_HUMAN               Reviewed;         777 AA.
AC   P04150; A0ZXF9; B0LPG8; D3DQF4; F5ATB7; P04151; Q53EP5; Q6N0A4;
DT   01-NOV-1986, integrated into UniProtKB/Swiss-Prot.
DT   01-NOV-1986, sequence version 1.
DT   11-DEC-2019, entry version 263.
DE   RecName: Full=Glucocorticoid receptor;
DE            Short=GR;
DE   AltName: Full=Nuclear receptor subfamily 3 group C member 1;
GN   Name=NR3C1; Synonyms=GRL;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS ALPHA AND BETA).
RC   TISSUE=Fibroblast;
RX   PubMed=2867473; DOI=10.1038/318635a0;
RA   Hollenberg S.M., Weinberger C., Ong E.S., Cerelli G., Oro A., Lebo R.,
RA   Thompson E.B., Rosenfeld M.G., Evans R.M.;
RT   "Primary structure and expression of a functional human glucocorticoid
RT   receptor cDNA.";
RL   Nature 318:635-641(1985).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] (ISOFORMS ALPHA AND BETA).
RX   PubMed=1707881;
RA   Encio I.J., Detera-Wadleigh S.D.;
RT   "The genomic structure of the human glucocorticoid receptor.";
RL   J. Biol. Chem. 266:7182-7188(1991).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   PubMed=20843780; DOI=10.1093/nar/gkq750;
RA   Wang W., Shen P., Thiyagarajan S., Lin S., Palm C., Horvath R.,
RA   Klopstock T., Cutler D., Pique L., Schrijver I., Davis R.W., Mindrinos M.,
RA   Speed T.P., Scharfe C.;
RT   "Identification of rare DNA variants in mitochondrial disorders with
RT   improved array-based sequencing.";
RL   Nucleic Acids Res. 39:44-58(2011).
```

Look at this UniProtKB entry (F1D8N4)

Note: the protein sequence found in this entry is 100 % identical to the canonical protein sequence of the UniProtKB/Swiss-Prot entry P04150

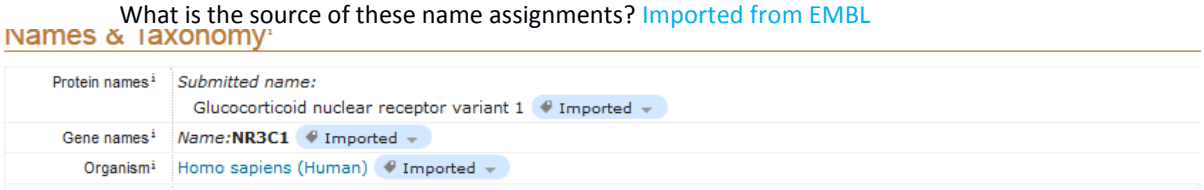
- 'Header section'

What is the status of this entry: reviewed by a biocurator or unreviewed?
What is the annotation score? 2/5
What is the evidence for the existence of the protein? Transcript level



- 'Names&Taxonomy' section

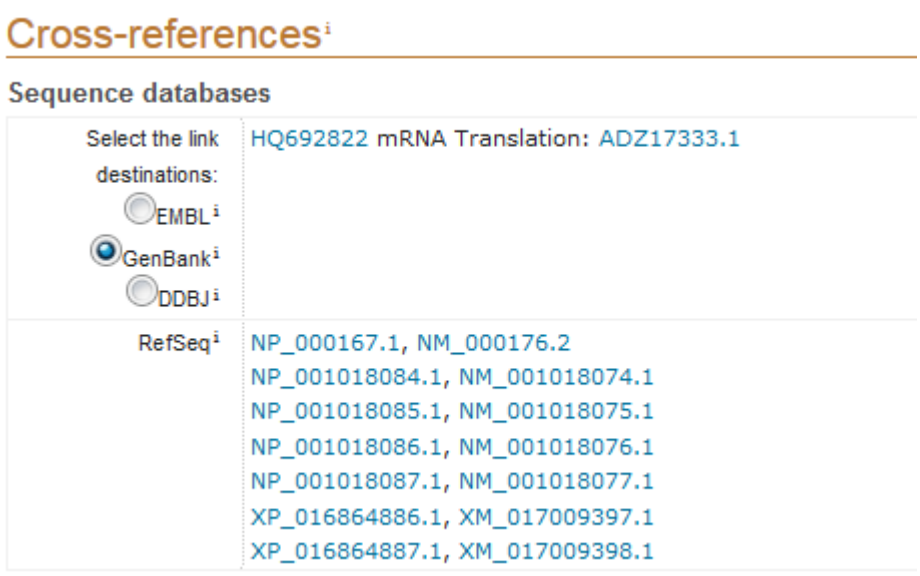
What are the name(s) of the gene and the name(s) of the protein?
What is the source of these name assignments? Imported from EMBL



- 'Cross references' section

Look (in GenBank) for the data available on the nucleic acid sequence.
Can you find a CoDing Sequence (CDS)? Where does the protein sequence come from?
Translation of the HQ692822 mRNA

How many RefSeq entries that have the same protein sequences?



Look at this entry (NP_000167.1)

Which database? (RefSeq, NCBI Reference Sequence) What is the main type of data? The protein sequence of glucocorticoid receptor (isoform alpha) (same as before) What is the length of the protein sequence? 777 aa

Important remark: the isoform name can differ between UniProtKB/Swiss-Prot and RefSeq !

GenPept

glucocorticoid receptor isoform alpha [Homo sapiens]

NCBI Reference Sequence: NP_000167.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_000167                777 aa            linear   PRI 09-FEB-2020
DEFINITION  glucocorticoid receptor isoform alpha [Homo sapiens].
```

Look for the annotation propagated from UniProtKB/Swiss-Prot?
Look for the annotation provided by 'CDD' (Conserved Domains and Protein Classification)

```
                     /calculated_mol_wt=85529
     Region          1..420
                     /region_name="Modulating"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            8
                     /site_type="phosphorylation"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphothreonine. {ECO:0000244|PubMed:23186163};
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            23
                     /site_type="methylation"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Omega-N-methylarginine.
                     {ECO:0000250|UniProtKB:P06537}; propagated from
                     UniProtKB/Swiss-Prot (P04150.1)"
     Region          26..401
                     /region_name="GCR"
                     /note="Glucocorticoid receptor; pfam02155"
                     /db_xref="CDD:308003"
```

Look at this UniProtKB entry (A0A0C5ESB7)

- 'Header section'

What is the status of this entry: reviewed by a biocurator or unreviewed?
What is the annotation score?
What is the evidence for the existence of the protein?

Protein **LexA repressor**

Gene **lexA**

Organism *Haemophilus influenzae*

Status Unreviewed - Annotation score: ●●●○○ - Protein inferred from homology[i]

- **'Names&Taxonomy'** section

What are the name(s) of the gene ?

What is the source of these name assignments?

Where does the EC (enzyme classification number) come from?)

Most of the information are imported from EMBL. Additional information come from automated annotation (UniRule, HAMAP)

Names & Taxonomy[i]

Protein names[i] *Recommended name:* **LexA repressor** UniRule annotation (EC:3.4.21.88 UniRule annotation)

Gene names[i] *Name:***lexA** UniRule annotation Imported

Synonyms:lexA_1 Imported, lexA_3 Imported

ORF Names:ABN48_00490 Imported, BLAS9_03375 Imported, BV055_00668 Imported, BV064_00575 Imported, BV163_00255 Imported, BV171_00384 Imported, BVZ80_00708 Imported, CH609_02840 Imported, CH620_08890 Imported, CH627_01025 Imported, FRB08_08395 Imported, FRB14_00155 Imported, NCTC11873_00263 Imported, NCTC11931_00467 Imported, NCTC13377_00255 Imported, NTHI1209_00187 Imported

Organism[i] Haemophilus influenzae Imported

- **'Function' section**

What is the function of the protein?

Where does the information come from?

Look at the UniRule (conditions and annotation)

Function[i]

Represses a number of genes involved in the response to DNA damage (SOS response), including recA and lexA. In the presence of single-stranded DNA, RecA interacts with LexA causing an autocatalytic cleavage which disrupts the DNA-binding part of LexA, leading to derepression of the SOS regulon and eventually DNA repair. UniRule annotation

Catalytic activity[i]

- Hydrolysis of Ala-|-Gly bond in repressor LexA. UniRule annotation SAAS annotation EC:3.4.21.88

Sites

Feature key	Position(s)	Description Actions	Graphical view	Length
Active site[i]	123	For autocatalytic cleavage activity UniRule annotation		1
Active site[i]	160	For autocatalytic cleavage activity UniRule annotation		1

If a protein meets these conditions... [i]

Common conditions

Matches HAMAP signature MF_00015
taxon = Bacteria
fragment ≠ the sequence is fragmented

Special conditions

taxon = Enterobacterales

taxon ≠ Enterobacterales

Subsequence at position 119 - 119 aligns to "S" in entry P0A7C2 (individually applies "For autocatalytic cleavage activity")
Subsequence at position 28 - 48 aligns to entry P0A7C2 (individually applies "H-T-H motif")
Subsequence at position 84 - 85 aligns to "A-G" in entry P0A7C2 (individually applies "Cleavage; by autolysis")
Subsequence at position 156 - 156 aligns to "K" in entry P0A7C2 (individually applies "For autocatalytic cleavage activity")

... then these annotations are applied [i]

Protein name[i]

Recommended name:
LexA repressor (EC:3.4.21.88)

Gene name[i]

*Name:*lexA

Subunit structure[i]

Homodimer.

Catalytic activity[i]

Hydrolysis of Ala-|-Gly bond in repressor LexA. EC:3.4.21.88

Function[i]

Represses a number of genes involved in the response to DNA damage (SOS response), including recA and lexA. Binds to the 16 bp palindromic sequence 5'-CTGTATATATATACAG-3'. In the presence of single-stranded DNA, RecA interacts with LexA causing an autocatalytic cleavage which disrupts the DNA-binding part of LexA, leading to derepression of the SOS regulon and eventually DNA repair.

Represses a number of genes involved in the response to DNA damage (SOS response), including recA and lexA. In the presence of single-stranded DNA, RecA interacts with LexA causing an autocatalytic

- 'Cross references' section

Look at the GenBank entry: JMQP01000002
What are the inferences available for the CDS of the LexA repressor?

```
CDS             37060..37683
                /gene="lexA"
                /locus_tag="NTHI1209_00187"
                /EC_number="3.4.21.88"
                /inference="ab initio prediction:Prodigal:2.60"
                /inference="similar to AA sequence:UniProtKB:P0A7C2"
                /codon_start=1
                /transl_table=11
                /product="LexA repressor"
                /protein_id="KIS34587.1"
                /translation="MRPLTARQQEVLDLLKRHLETTGMPPTRAEISRELGFKSANAAE
                EHLKALSRKGAIEIIPGASRGIRILDNSSNDEFDGLPLVGRVAAGEPILAEQHIEATY
                RVDADMFKPQADFLLKVYGLSMKNVGILDGDLLAVHSTKDVRNGQIVVARIEDEVTVK
                RLEKKGSIIYLHAENEEFDPIVVNLEEQKNFEIEGIAVGIIRNNAWM"
```

Look at the annotation of the corresponding original entry in RefSeq:
https://www.ncbi.nlm.nih.gov/protein/NP_438908.2

GenPept ▾ Send to: ▾

LexA repressor [Haemophilus influenzae Rd KW20]

NCBI Reference Sequence: NP_438908.2
Identical Proteins FASTA Graphics

Go to: ⊙

```
LOCUS       NP_438908               207 aa            linear   CON 02-AUG-2016
DEFINITION  LexA repressor [Haemophilus influenzae Rd KW20].
ACCESSION   NP_438908
VERSION     NP_438908.2
DBLINK      BioProject: PRJNA57771
            Assembly: GCF_000027305.1
DBSOURCE    REFSEQ: accession NC_000907.1
KEYWORDS    RefSeq.
SOURCE      Haemophilus influenzae Rd KW20
  ORGANISM  Haemophilus influenzae Rd KW20
            Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales;
            Pasteurellaceae; Haemophilus.
REFERENCE   1  (residues 1 to 207)
  AUTHORS   Gmuender,H., Kuratli,K., Di Padova,K., Gray,C.P., Keck,W. and
            Evers,S.
  TITLE     Gene expression changes triggered by exposure of Haemophilus
            influenzae to novobiocin or ciprofloxacin: combined transcription
            and translation analysis
  JOURNAL   Genome Res. 11 (1), 28-42 (2001)
   PUBMED   11156613
REFERENCE   2  (residues 1 to 207)
  AUTHORS   Langen,H., Takacs,B., Evers,S., Berndt,P., Lahm,H.W., Wipf,B.,
            Gray,C. and Fountoulakis,M.
  TITLE     Two-dimensional map of the proteome of Haemophilus influenzae
  JOURNAL   Electrophoresis 21 (2), 411-429 (2000)
   PUBMED   10675023
REFERENCE   3  (residues 1 to 207)
  AUTHORS   Fountoulakis,M., Juranville,J.F., Roder,D., Evers,S., Berndt,P. and
            Langen,H.
  TITLE     Reference map of the low molecular mass proteins of Haemophilus
            influenzae
  JOURNAL   Electrophoresis 19 (10), 1819-1827 (1998)
   PUBMED   9719565
```

Look at the annotation of the corresponding 'reannotated' entry in RefSeq (MULTISPECIES):
https://www.ncbi.nlm.nih.gov/protein/WP_005648504.1

GenPept

This record is a non-redundant protein sequence. Please read more here.

MULTISPECIES: repressor LexA [Haemophilus]

NCBI Reference Sequence: WP_005648504.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_005648504            207 aa            linear   BCT 16-JUN-2019
DEFINITION  MULTISPECIES: repressor LexA [Haemophilus].
ACCESSION   WP_005648504
VERSION     WP_005648504.1
KEYWORDS    RefSeq.
SOURCE      Haemophilus
  ORGANISM  Haemophilus
            Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales;
            Pasteurellaceae.
REFERENCE   1  (residues 1 to 207)
  AUTHORS   Winterling,K.W., Chafin,D., Hayes,J.J., Sun,J., Levine,A.S.,
            Yasbin,R.E. and Woodgate,R.
  TITLE     The Bacillus subtilis DinR binding site: redefinition of the
            consensus sequence
  JOURNAL   J. Bacteriol. 180 (8), 2201-2211 (1998)
   PUBMED   9555905
REFERENCE   2  (residues 1 to 207)
  AUTHORS   Harmon,F.G., Rehrauer,W.M. and Kowalczykowski,S.C.
  TITLE     Interaction of Escherichia coli RecA protein with LexA repressor.
            II. Inhibition of DNA strand exchange by the uncleavable LexA S119A
            repressor argues that recombination and SOS induction are
            competitive processes
  JOURNAL   J. Biol. Chem. 271 (39), 23874-23883 (1996)
   PUBMED   8798618
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
```

Look at the 'identical protein' sequences in RefSeq:
https://www.ncbi.nlm.nih.gov/ipg/WP_005648504.1

Identical Protein Groups Send to:

repressor LexA

GenPept FASTA Graphics BLAST

Name: repressor LexA
RefSeq Selected Product: WP_005648504.1, 207 amino acids
Assembly Accessions: 1132
Protein Accessions: 483
CDS Regions: 1150
Total Rows: 1153

<<First <Prev Page 1 of 24 (50 CDSs per page) Next > Last>>

Source (All)	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Assembly
RefSeq	NC_007146.2 876448-877071 (+)	WP_005648504.1	repressor LexA	Haemophilus influenzae 86-028NP	86-028NP	GCF_000012185.1
RefSeq	NC_009566.1 1641809-1642432 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae PittEE	PittEE	GCF_000016465.1
RefSeq	NC_009567.1 1342147-1342770 (+)	WP_005648504.1	repressor LexA	Haemophilus influenzae PittGG	PittGG	GCF_000016485.1
RefSeq	NC_014920.1 1465418-1466041 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae F3031	F3031	GCF_000197875.1
RefSeq	NC_014922.1 390129-390752 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae F3047	F3047	GCF_000200475.1
RefSeq	NC_016809.1 915653-916276 (+)	WP_005648504.1	repressor LexA	Haemophilus influenzae 10810	10810	GCF_000210875.1
RefSeq	NC_017451.1 1678112-1678735 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae R2866	R2866	GCF_000165525.1
RefSeq	NC_017452.1 1640394-1641017 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae R2846	R2846	GCF_000165575.1
RefSeq	NC_022356.1 191450-192073 (+)	WP_005648504.1	repressor LexA	Haemophilus influenzae KR494	KR494	GCF_000465255.1
RefSeq	NZ_AAZD01000001.1 567018-567641 (-)	WP_005648504.1	repressor LexA	Haemophilus influenzae 22.1-21	22.1-21	GCF_000169735.1

Look at the corresponding UniRef100 at UniProtKB: https://www.uniprot.org/uniref/UniRef100_Q4QME9

Members

Expand cluster to 90% or 50% identity.
1 to **12** of **12** Show 25

Cluster members	Entry name		Protein names	Organisms
Q4QME9	**LEXA_HAEI8**		**LexA repressor**	**Haemophilus influenzae (strain 86-028NP)**
A5UDX6	LEXA_HAEIE		LexA repressor	Haemophilus influenzae (strain PittEE)
A5UHQ3	LEXA_HAEIG		LexA repressor	Haemophilus influenzae (strain PittGG)
A0A5C2QZX1	A0A5C2QZX1_HAEIF		LexA repressor	Haemophilus influenzae biotype aegyptius
A0A0C5ESB7	A0A0C5ESB7_HAEIF		LexA repressor	Haemophilus influenzae
A4MWY4	A4MWY4_HAEIF		LexA repressor	Haemophilus influenzae 22.1-21
A4NWV7	A4NWV7_HAEIF		LexA repressor	Haemophilus influenzae 22.4-21
E7A6E5	E7A6E5_HAEIF		LexA repressor	Haemophilus influenzae F3031
A0A448MDL7	A0A448MDL7_HAEAE		LexA repressor	Haemophilus aegyptius
A4N2Y6	A4N2Y6_HAEIF		LexA repressor	Haemophilus influenzae R3021
A0A0E1SS24	A0A0E1SS24_HAEIF		LexA repressor	Haemophilus influenzae PittII
UPI000D017131			repressor LexA	Haemophilus influenzae

Look at this entry (https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512)

- Which database? What is the main type of data?

  GenBank; Genome of coronavirus COVID-19

GenBank ▾ Send to:

Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

FASTA Graphics

```
LOCUS       NC_045512              29903 bp ss-RNA     linear   VRL 28-JAN-2020
DEFINITION  Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete
            genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae;
            Orthocoronavirinae; Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G.,
            Tao,Z.-W., Tian,J.-H., Pei,Y.-Y., Yuan,M.L., Zhang,Y.-L.,
            Dai,F.-H., Liu,Y., Wang,Q.-M., Zheng,J.-J., Xu,L., Holmes,E.C. and
            Zhang,Y.-Z.
  TITLE     A novel coronavirus associated with a respiratory disease in Wuhan
            of Hubei province, China
  JOURNAL   Unpublished
```

- How many predicted proteins? (follow the link to 'Protein')

12 proteins

https://www.ncbi.nlm.nih.gov/protein?LinkName=nuccore_protein&from_uid=1798174254

- The corresponding proteins are also available on UniProtKB ftp: ftp://ftp.uniprot.org/pub/databases/uniprot/pre_release/

Compare the annotation of the protein of 7096 aa provided by RefSeq (https://www.ncbi.nlm.nih.gov/protein/YP_009724389.1) and by UniProtKB (ftp://ftp.uniprot.org/pub/databases/uniprot/pre_release/coronavirus.dat -> R1AB_WCPV)

- How many proteins for this virus in UniProtKB ? 14
  ftp://ftp.uniprot.org/pub/databases/uniprot/pre%5Frelease/coronavirus.fasta

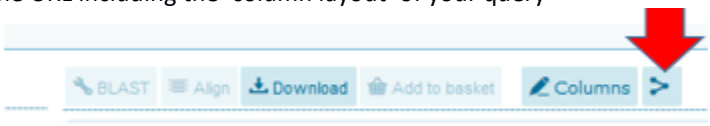
More information of the annotation of these proteins this afternoon….

Discover the UniProt Search tool

(a) Look for the gene PRSS1 in UniProtKB
Restrict 'PRSS1' to exact gene name.

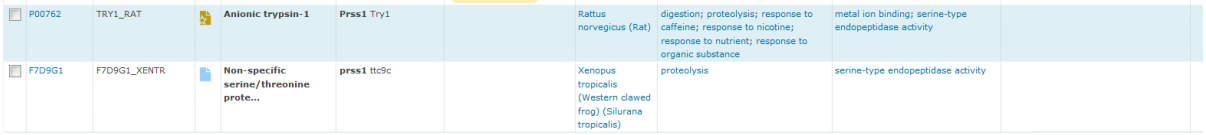
(b) Customize your results
Configure your column layout by adding and removing columns, until you have columns for 'Protein names', 'Reviewed/Unreviewed', 'Organism', '3D' (PDB), 'Gene Ontology (Biological process / Molecular function)', 'Function' (Function CC).

Look at the URL including the 'column layout' of your query



https://www.uniprot.org/uniprot/?query=gene%3A%20PRSS1%20&columns=id%2Centry%20name%2Creviewed%2Cprotein%20names%2Cgenes%2Ccomment(FUNCTION)%2Corganism%2Cgo(biological%20process)%2Cgo(molecular%20function)%2Cdatabase(PDB)&sort=score

(c) Compare the annotations of reviewed and unreviewed entries.

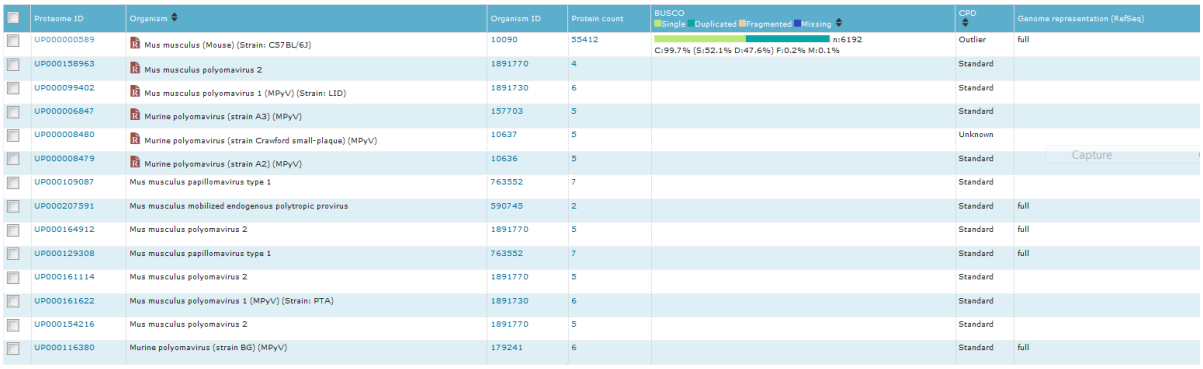


Proteomes

What is the UniProt release number you are working with?

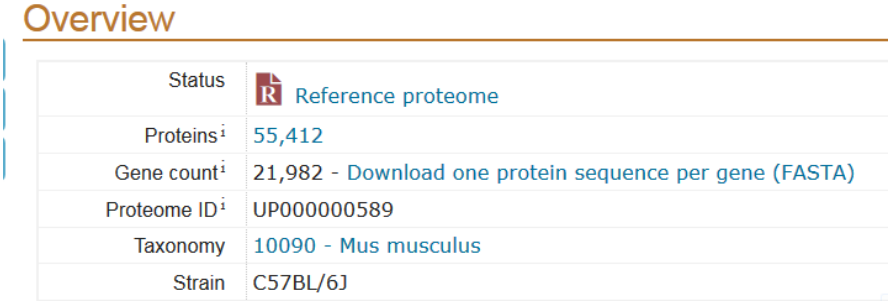
The mouse proteome at UniProtKB
Query UniProt Proteomes

What is the Proteome ID of the reference Proteome? UP000000589
To which strain does it correspond? (Strain: C57BL/6J)

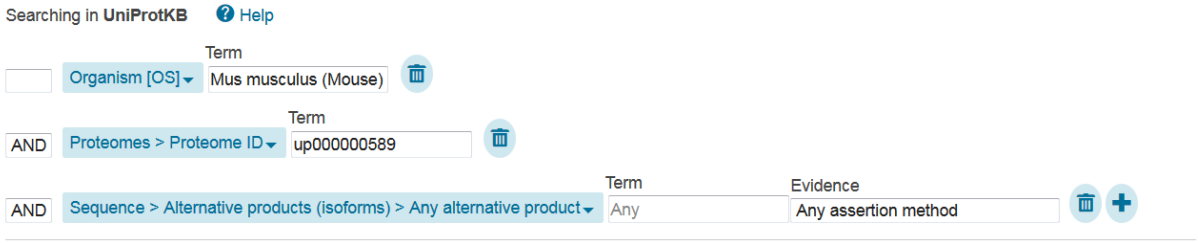


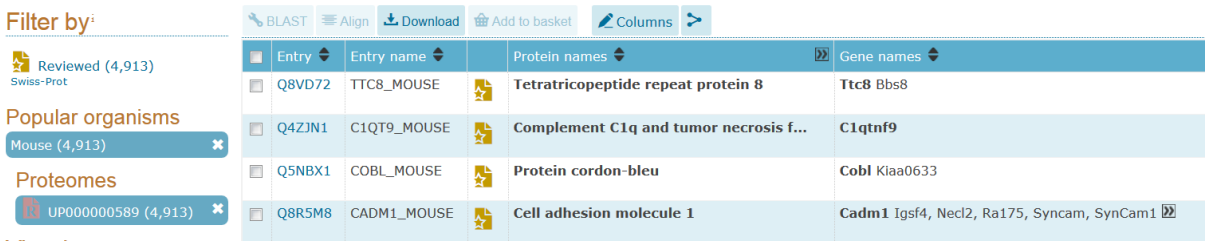
Proteome ID	Organism	Organism ID	Protein count	BUSCO Single Duplicated Fragmented Missing	CPD	Genome representation (RefSeq)
UP000000589	Mus musculus (Mouse) (Strain: C57BL/6J)	10090	55412	n:6192 C:99.7% (S:52.1% D:47.6%) F:0.2% M:0.1%	Outlier	full
UP000158963	Mus musculus polyomavirus 2	1891770	4		Standard	
UP000099402	Mus musculus polyomavirus 1 (MPyV) (Strain: LID)	1891730	6		Standard	
UP000006847	Murine polyomavirus (strain A3) (MPyV)	157703	5		Standard	
UP000008480	Murine polyomavirus (strain Crawford small-plaque) (MPyV)	10637	5		Unknown	
UP000008479	Murine polyomavirus (strain A2) (MPyV)	10636	5		Standard	
UP000109087	Mus musculus papillomavirus type 1	763552	7		Standard	
UP000207591	Mus musculus mobilized endogenous polytropic provirus	590745	2		Standard	full
UP000164912	Mus musculus polyomavirus 2	1891770	5		Standard	full
UP000129308	Mus musculus papillomavirus type 1	763552	7		Standard	full
UP000161114	Mus musculus polyomavirus 2	1891770	5		Standard	
UP000161622	Mus musculus polyomavirus 1 (MPyV) (Strain: PTA)	1891730	6		Standard	
UP000154216	Mus musculus polyomavirus 2	1891770	5		Standard	
UP000116380	Murine polyomavirus (strain BG) (MPyV)	179241	6		Standard	full

- How many estimated gene number?
- How many estimated protein sequences?

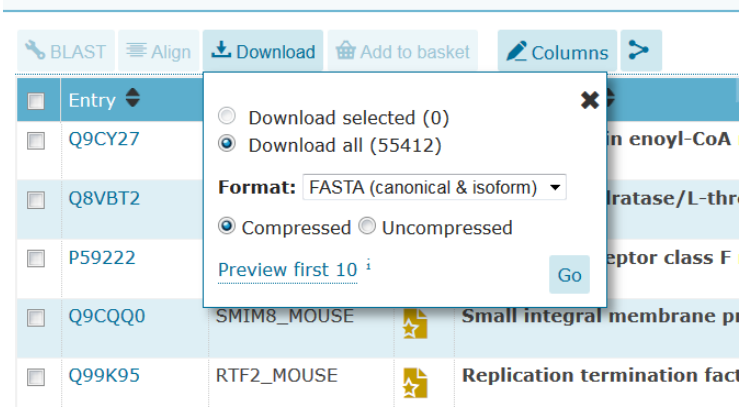


- How many records describe more than one protein sequence? 4913
  Hint: how many records with 'sequence / alternative products'





- ...download the mouse proteome sequences in fasta format, including the sequences of additional isoforms = FASTA (canonical & isoform).



2. The following SPARQL query allows to know the number of alternative sequences ('additional isoforms') which are have are found in UniProtKB/Swiss-Prot (human proteome).

Go to : https://sparql.uniprot.org/

```
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX proteome:<http://purl.uniprot.org/proteomes/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT (COUNT(DISTINCT ?sequence) AS ?allIsoforms)
WHERE
{
   ?protein up:reviewed true .
    ?protein up:organism taxon:9606 .
   ?protein up:sequence ?sequence .
   ?protein up:proteome/^skos:narrower proteome:UP000005640 .
}
```

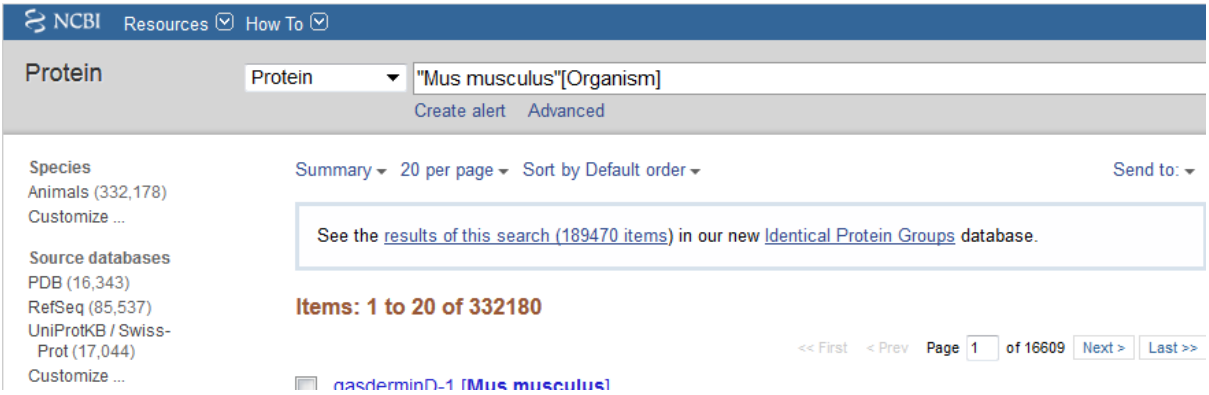
Adapt the SPARQL query in order to know the number of 'additional isoforms' found in the mouse proteome in UniProtKB/Swiss-Prot.

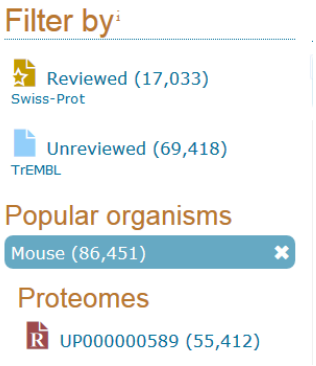
```
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX proteome:<http://purl.uniprot.org/proteomes/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT (COUNT(DISTINCT ?sequence) AS ?allIsoforms)
WHERE
{
   ?protein up:reviewed true .
    ?protein up:organism taxon:10090 .
   ?protein up:sequence ?sequence .
   ?protein up:proteome/^skos:narrower proteome:UP000000589 .
}
```

➔ 25’311 additional protein sequences
➔ 55’412 + 25’311 protein sequences in the UniProtKB mouse proteome.

3. The mouse proteome at NCBI protein

How many protein sequences for Mus musculus @NCBI protein?
How many protein sequences in RefSeq?
Why this difference compared to UniProtKB?





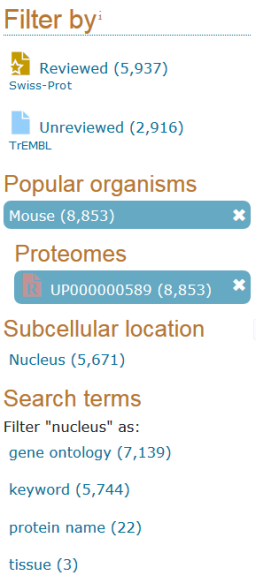
UniProtKB/Swiss-Prot: one entry per gene. Several protein sequences per entry (alternative products / ‘additional isoforms);
UniProtKB/TrEMBL: one entry per protein sequence
UniProtKB proteome: set of sequences mapped to the mouse genome
RefSeq: mainly one entry per mRNA.

4. The mouse nuclear proteome at UniProtKB

- How many mouse proteins are found in the ‘nucleus’ (full text search)

Query: nucleus AND organism:"Mus musculus (Mouse) [10090]" AND proteome:up000000589

- In which field(s) do you find the annotation (look at the filters, column on the left)?



- Compare the number of entries with have the location in the nucleus deduced from GO annotation and those annotated with 'Subcellular location Nucleus'

The keywords and Subcellular location are mainly annotated manually by the UniProt biocurators. This type of annotation represents a 'filtered summary' of what is known.
The Gene ontology is annotated by the MGI team (or other model databases) and is imported in UniProt: the GO annotation gives an overview of all what is known or has been found in the publications. Beware: GO is an ontology, each term can have parents and children.

Further information:
'What are proteomes?'
'What are reference proteomes?'
'How to retrieve sets of protein sequences?'

N-glycosylation: compare different datasets

N-linked glycosylation is the most important form of post-translational modification for proteins synthesized and folded in the Endoplasmic Reticulum.
Imagine you have been working on a mouse protein family where 20% of all members are N-glycosylated. Try to find out whether your protein family is glycosylated significantly more frequently than other proteins.
Hints:
In the mouse proteome:

- What is the percentage of N-glycosylated entries?
- What is the percentage of reviewed entries?
- What is the percentage of N-glycosylated entries in reviewed and unreviewed entries, respectively?
- Try to find out whether your protein family is glycosylated significantly more frequently than other proteins.

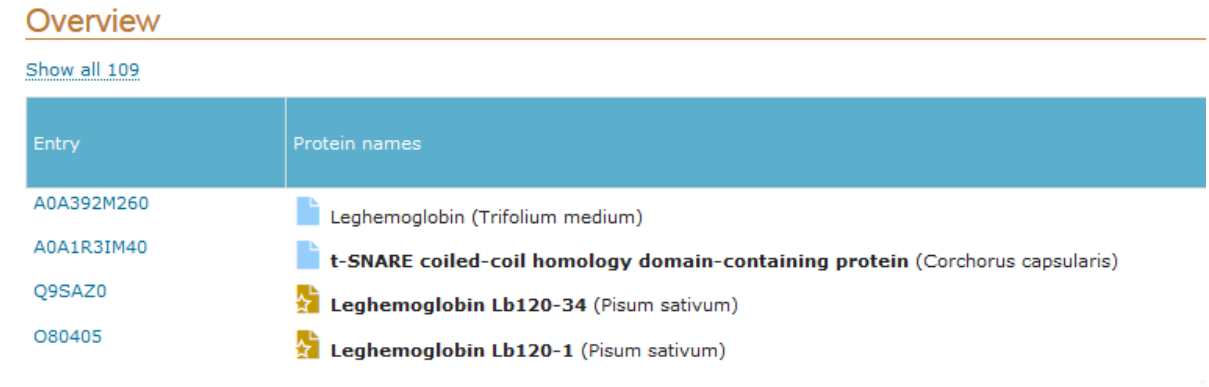
Set of mouse proteins with N-glycosylation – Feb 2020

Reviewed (17,027) Swiss-Prot Unreviewed (69,420) TrEMBL Popular organisms Mouse (86,447) Proteomes UP000000589 (55,408)	Reviewed (3,710) Swiss-Prot Unreviewed (43) TrEMBL Popular organisms Mouse (3,753) Proteomes UP000000589 (3,730)	21.7 % 0.06 % 6.7 %
organism:"Mus musculus (Mouse) [10090]"	annotation:(type:carbohyd "n linked glcnac ellipsis") AND organism:"Mus musculus (Mouse) [10090]"	

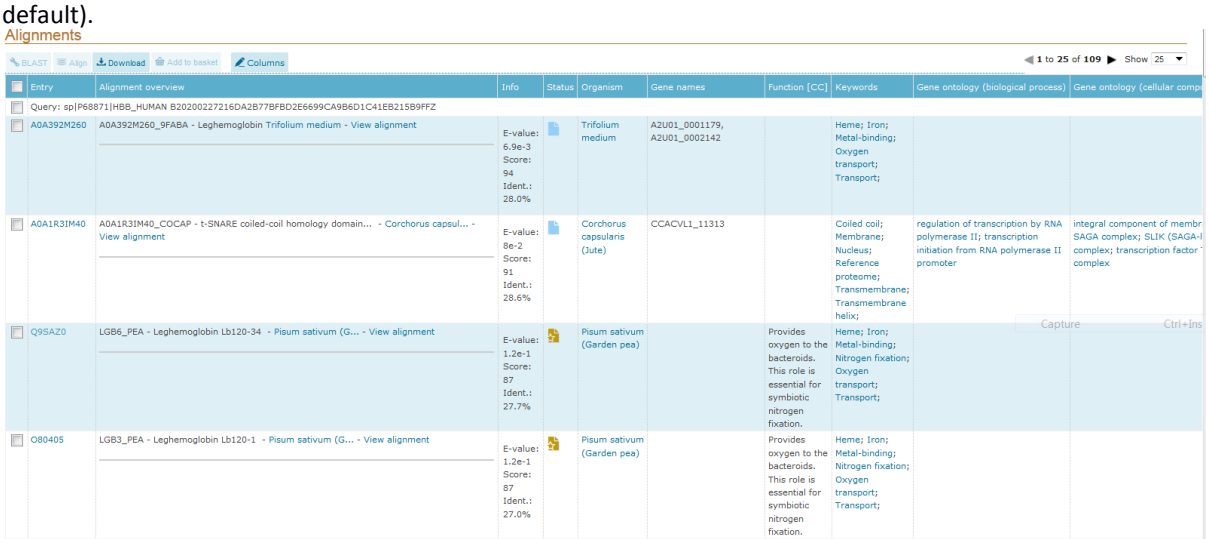
Relevant UniProt help pages:
- query syntax
- Glycosylation

Discover the UniProt BLAST tool

Look for plant protein sequences similar to human hemoglobin (HBB):
BLAST the human hemoglobin HBB sequence against 'Plant' sequences in UniProtKB (use Advanced BLAST).

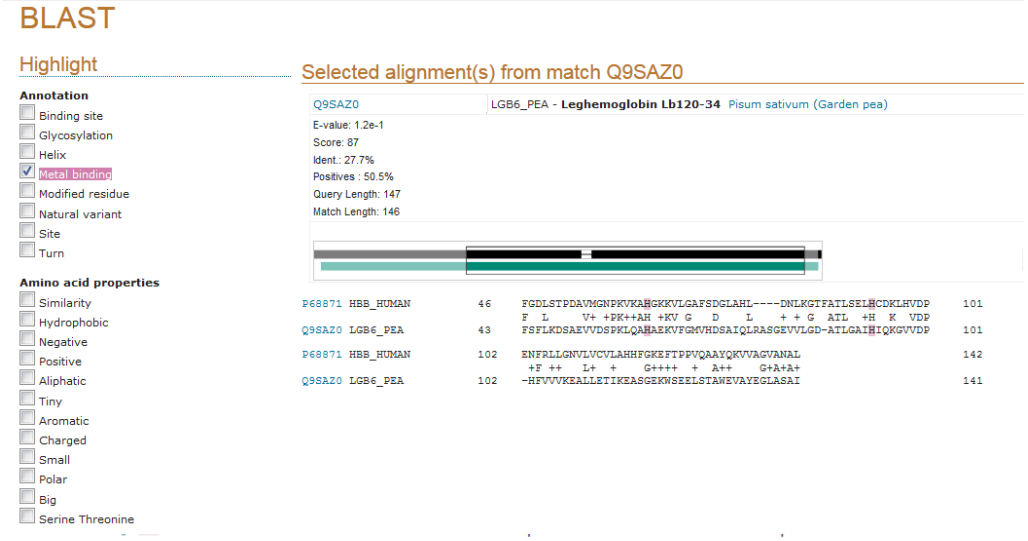


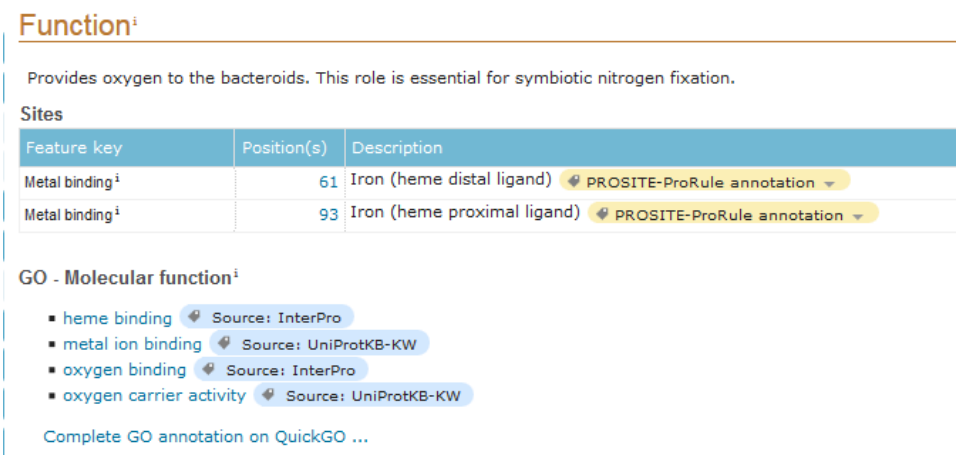
- *Customize your 'BLAST results'*
Add columns for protein names, gene names, function, keywords, gene ontology (some are already there by default).



- For the first matching UniProtKB/Swiss-Prot entry, open the pairwise alignment

Look at the conservation of the iron binding sites ('Metal binding' in the “Highlight” options on the left).





Discover the UniProt ID mapping tool

Which database do these identifiers correspond to?

NP_001018084 NP_001018085 NP_001018086 NP_001191191 NP_001191192 NP_001191193 XP_005268476 XP_005268477 XP_016864886 XP_016864887

Find the corresponding UniProtKB entries, using UniProt's ID mapping tool.
Do a multiple alignment of the UniProtKB entries. How many differences? Why these differences?

Retrieve/ID mapping

How to use Retrieve/ID mapping tool

Enter or upload a list of identifiers to do one of the following:

Retrieve the corresponding UniProt entries to download them or work with them on this website.
Convert identifiers which are of a different type to UniProt identifiers or vice versa and download the identifier lists.

1. Provide your identifiers

```
NP_001018084 NP_001018085 NP_001018086 NP_001191191 NP_001191192 NP_001191193 XP_005268476
XP_005268477 XP_016864886 XP_016864887
```

OR upload your own file: Browse... No file selected.

Run in a new window.

2. Select options

From: RefSeq Protein

To: UniProtKB

BLAST | Align | Download | Add to basket | Columns | 1 to 3 of 3 | Show 25

Your list:...ECB7B6	Isoform map:...ECB7B6	Entry	Entry name		Protein names	Gene names	Organism	Length
XP_005268476, XP_005268477		E5KQF6	E5KQF6_HUMAN		Glucocorticoid receptor	NR3C1 hCG_37601	Homo sapiens (Human)	778
NP_001018084, NP_001018085, NP_001018086, XP_016864886, XP_016864887		F1D8N4	F1D8N4_HUMAN		Glucocorticoid nuclear receptor var...	NR3C1	Homo sapiens (Human)	777
NP_001018084, NP_001018085, NP_001018086, NP_001191191, NP_001191192, NP_001191193, XP_005268476, XP_005268477, XP_016864886, XP_016864887	NP_001018084 -> P04150-1NP_001018086 -> P04150-1NP_001018085 -> P04150-1NP_001191193 -> P04150-16NP_001191192 -> P04150-15NP_001191191 -> P04150-14	P04150	GCR_HUMAN		Glucocorticoid receptor	NR3C1 GRL	Homo sapiens (Human)	777

Job status: COMPLETED

```
E5KQF6_HUMAN      1  MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR   60
F1D8N4_HUMAN      1  MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR   60
P04150_GCR_HUMAN  1  MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR   60
                     ************************************************************

E5KQF6_E5KQF6_HUMAN  61  LLVDFPKGSVSNAQQPDLSKAVSLSMGLYMGETETKVMGNDLGFPQQGQISLSSGETDLK  120
F1D8N4_F1D8N4_HUMAN  61  LLVDFPKGSVSNAQQPDLSKAVSLSMGLYMGETETKVMGNDLGFPQQGQISLSSGETDLK  120
P04150_GCR_HUMAN     61  LLVDFPKGSVSNAQQPDLSKAVSLSMGLYMGETETKVMGNDLGFPQQGQISLSSGETDLK  120
                         ************************************************************

E5KQF6_E5KQF6_HUMAN  121  LLEESIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQQHLKGQTGTNGGN  180
F1D8N4_F1D8N4_HUMAN  121  LLEESIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQQHLKGQTGTNGGN  180
P04150_GCR_HUMAN     121  LLEESIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQQHLKGQTGTNGGN  180
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  181  VKLYTTDQSTFDILQDLEFSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN  240
F1D8N4_F1D8N4_HUMAN  181  VKLYTTDQSTFDILQDLEFSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN  240
P04150_GCR_HUMAN     181  VKLYTTDQSTFDILQDLEFSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN  240
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  241  SNEDCKPLILPDTKPKIKDNGDLVLSSPSNVTLPQVKTEKEDFIELCTPGVIKQEKLGTV  300
F1D8N4_F1D8N4_HUMAN  241  SNEDCKPLILPDTKPKIKDNGDLVLSSPSNVTLPQVKTEKEDFIELCTPGVIKQEKLGTV  300
P04150_GCR_HUMAN     241  SNEDCKPLILPDTKPKIKDNGDLVLSSPSNVTLPQVKTEKEDFIELCTPGVIKQEKLGTV  300
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  301  YCQASFPGANIIGNKMSAISVHGVSTSGGQMYHYDMNTASLSQQQDQKPIFNVIPPIPVG  360
F1D8N4_F1D8N4_HUMAN  301  YCQASFPGANIIGNKMSAISVHGVSTSGGQMYHYDMNTASLSQQQDQKPIFNVIPPIPVG  360
P04150_GCR_HUMAN     301  YCQASFPGANIIGNKMSAISVHGVSTSGGQMYHYDMNTASLSQQQDQKPIFNVIPPIPVG  360
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  361  SENWNRCQGSGDDNLTSLGTLNFPGRTVFSNGYSSPSMRPDVSSPPSSSSTATTGPPPKL  420
F1D8N4_F1D8N4_HUMAN  361  SENWNRCQGSGDDNLTSLGTLNFPGRTVFSNGYSSPSMRPDVSSPPSSSSTATTGPPPKL  420
P04150_GCR_HUMAN     361  SENWNRCQGSGDDNLTSLGTLNFPGRTVFSNGYSSPSMRPDVSSPPSSSSTATTGPPPKL  420
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  421  CLVCSDEASGCHYGVLTCGSCKVFFKRAVEGRQHNYLCAGRNDCIIDKIRRKNCPACRYR  480
F1D8N4_F1D8N4_HUMAN  421  CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGRNDCIIDKIRRKNCPACRYR  479
P04150_GCR_HUMAN     421  CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGRNDCIIDKIRRKNCPACRYR  479
                          ********************************* ********************************

E5KQF6_E5KQF6_HUMAN  481  KCLQAGMNLEARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQLTPTLVSLLEVI  540
F1D8N4_F1D8N4_HUMAN  480  KCLQAGMNLEARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQLTPTLVSLLEVI  539
P04150_GCR_HUMAN     480  KCLQAGMNLEARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQLTPTLVSLLEVI  539
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  541  EPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLLQYS  600
F1D8N4_F1D8N4_HUMAN  540  EPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLLQYS  599
P04150_GCR_HUMAN     540  EPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLLQYS  599
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  601  WMFLMAFALGWRSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSELHRLQVS  660
F1D8N4_F1D8N4_HUMAN  600  WMFLMAFALGWRSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSELHRLQVS  659
P04150_GCR_HUMAN     600  WMFLMAFALGWRSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSELHRLQVS  659
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  661  YEEYLCMKTLLLLSSVPKDGLKSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLT  720
F1D8N4_F1D8N4_HUMAN  660  YEEYLCMKTLLLLSSVPKDGLKSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLT  719
P04150_GCR_HUMAN     660  YEEYLCMKTLLLLSSVPKDGLKSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLT  719
                          ************************************************************

E5KQF6_E5KQF6_HUMAN  721  KLLDSMHEVVENLLNYCFQTFLDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK  778
F1D8N4_F1D8N4_HUMAN  720  KLLDSMHEVVENLLNYCFQTFLDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK  777
P04150_GCR_HUMAN     720  KLLDSMHEVVENLLNYCFQTFLDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK  777
                          ************************************************************
```

Links to

- **User manual**
- **FAQ**
- **Documentation/Help**

Elisabeth Gasteiger, Marie-Claude Blatter
SIB Swiss Institute of Bioinformatics, March 2020