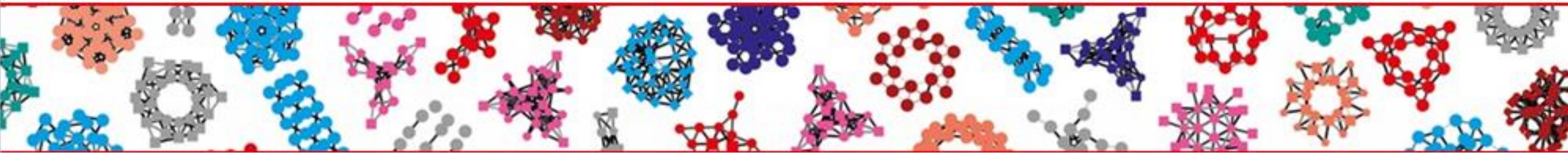


UniProtKB: manual and automated protein sequence annotation pipelines

WELCOME !

http://education.expasy.org/cours/SIB_UniProtKB_2018/

Basel, November 2018

Elisabeth.Gasteiger@sib.swiss

Marie-Claude.Blatter@sib.swiss

Swiss-Prot group
SIB Swiss Institute of Bioinformatics

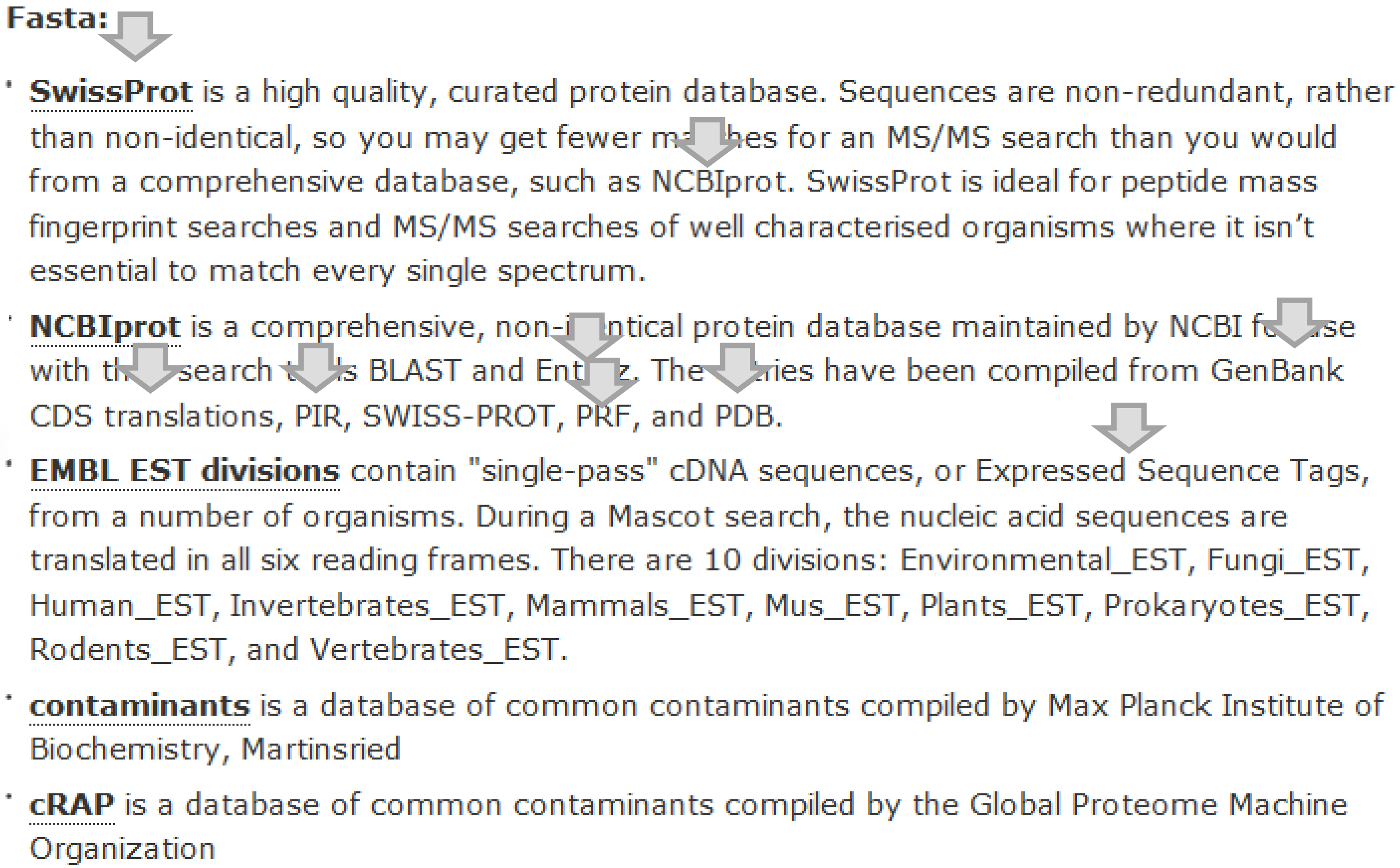
The **goals of this course (theory + practicals)** are to give some basic theoretical and practical knowledge on protein sequence databases with a **focus on UniProtKB, on the different manual and automated annotation pipelines and on the optimum use of UniProt.**

We will discuss questions such as:

- Where do the protein sequences come from?
- What are the differences between the major protein sequence databases?
- What are the manual and automated gene / protein annotation pipelines?
- How to assess protein sequence accuracy and annotation quality?
- How to extract biological knowledge from a Blast result, a gene list or a multiple alignment?

- This afternoon: focus on automated annotation pipelines http://www.**matrixscience**.com/search_intro.html

Protein sequences



Fasta:

- **SwissProt** is a high quality, curated protein database. Sequences are non-redundant, rather than non-identical, so you may get fewer matches for an MS/MS search than you would from a comprehensive database, such as NCBIprot. SwissProt is ideal for peptide mass fingerprint searches and MS/MS searches of well characterised organisms where it isn't essential to match every single spectrum.
- **NCBIprot** is a comprehensive, non-identical protein database maintained by NCBI for use with the search tools BLAST and Entrez. The entries have been compiled from GenBank CDS translations, PIR, SWISS-PROT, PRF, and PDB.
- **EMBL EST divisions** contain "single-pass" cDNA sequences, or Expressed Sequence Tags, from a number of organisms. During a Mascot search, the nucleic acid sequences are translated in all six reading frames. There are 10 divisions: Environmental_EST, Fungi_EST, Human_EST, Invertebrates_EST, Mammals_EST, Mus_EST, Plants_EST, Prokaryotes_EST, Rodents_EST, and Vertebrates_EST.
- **contaminants** is a database of common contaminants compiled by Max Planck Institute of Biochemistry, Martinsried
- **cRAP** is a database of common contaminants compiled by the Global Proteome Machine Organization

IPI
RefSeq
GenPept
…

Where do the protein sequences come from ?
What's about the sequence accuracy ?

Biological knowledge

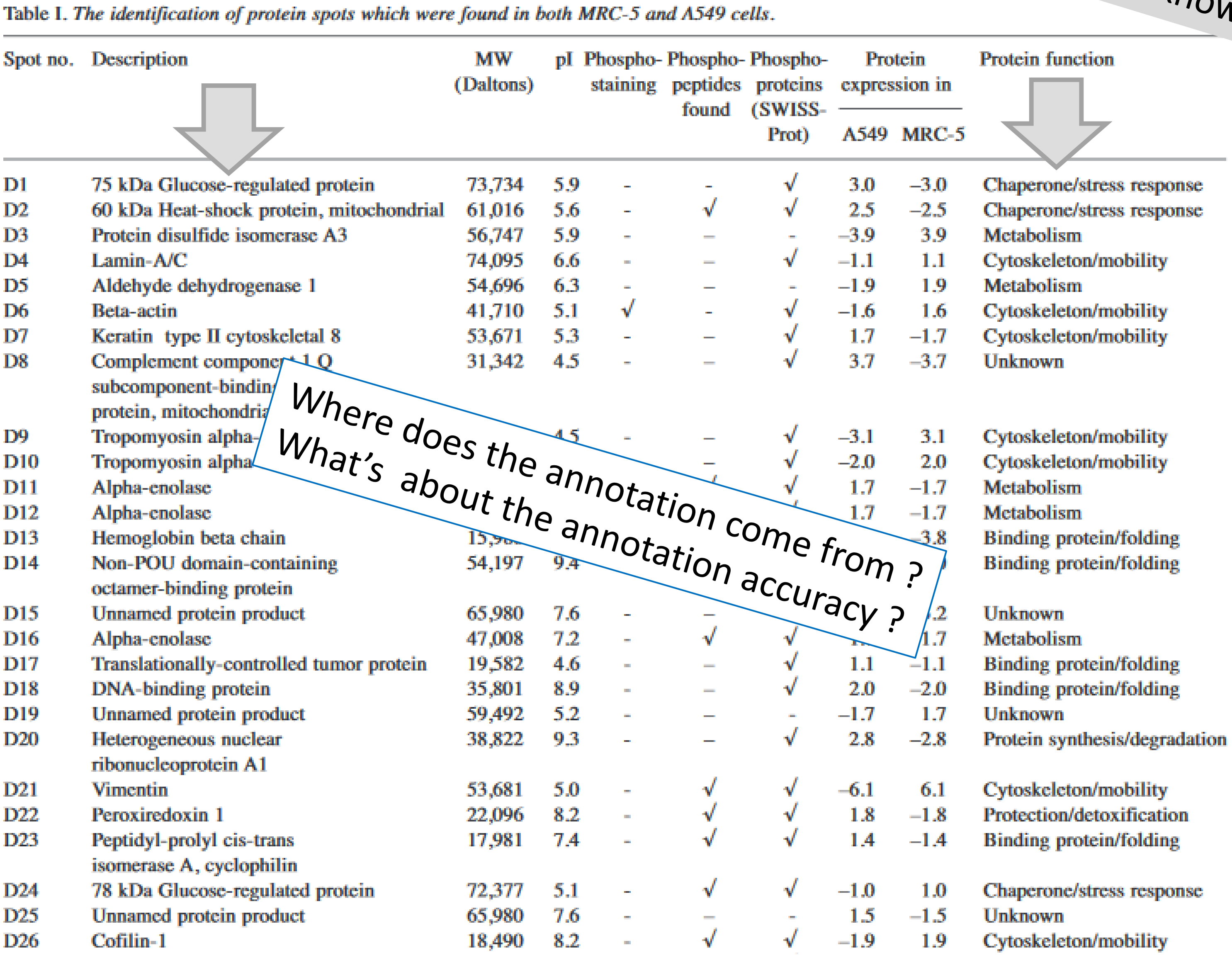


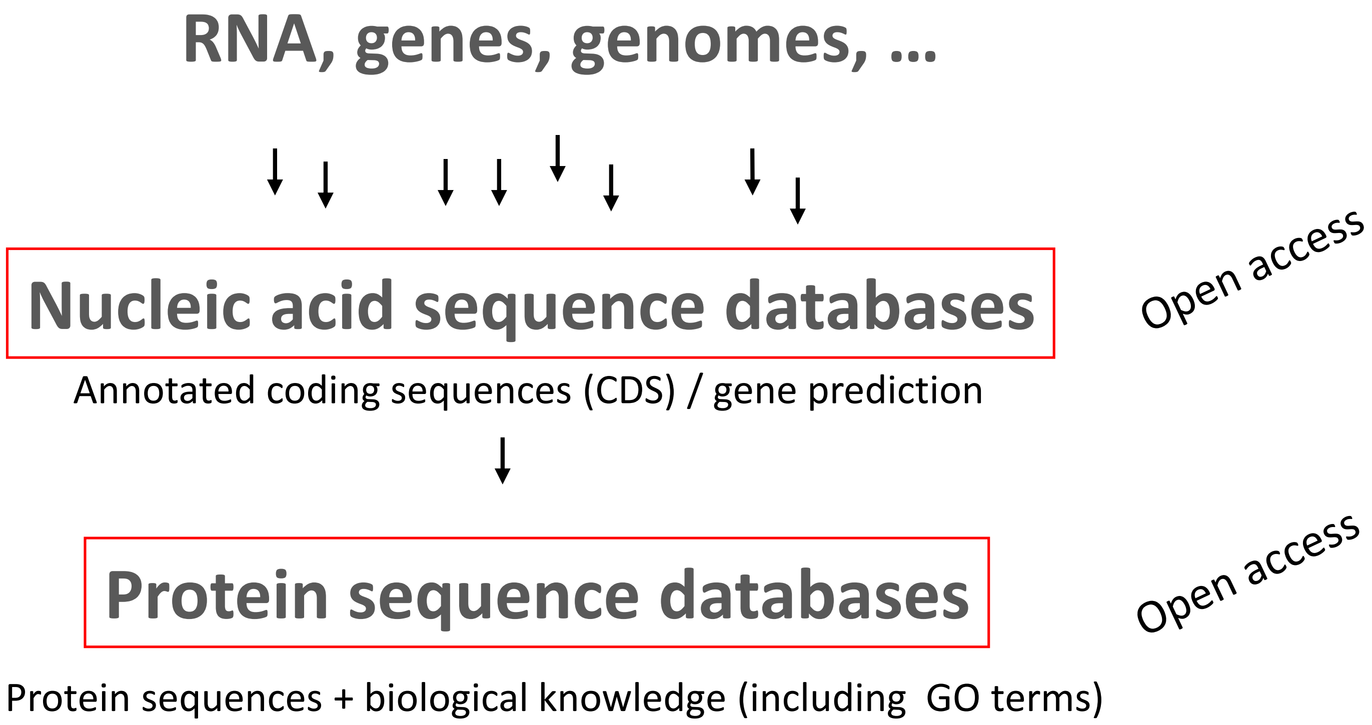
Table I. *The identification of protein spots which were found in both MRC-5 and A549 cells.*

Spot no.	Description	MW (Daltons)	pI	Phospho-staining	Phospho-peptides found	Phospho-proteins (SWISS-Prot)	Protein expression in		Protein function
							A549	MRC-5	
D1	75 kDa Glucose-regulated protein	73,734	5.9	-	-	√	3.0	–3.0	Chaperone/stress response
D2	60 kDa Heat-shock protein, mitochondrial	61,016	5.6	-	√	√	2.5	–2.5	Chaperone/stress response
D3	Protein disulfide isomerase A3	56,747	5.9	-	–	-	–3.9	3.9	Metabolism
D4	Lamin-A/C	74,095	6.6	-	–	√	–1.1	1.1	Cytoskeleton/mobility
D5	Aldehyde dehydrogenase 1	54,696	6.3	-	–	-	–1.9	1.9	Metabolism
D6	Beta-actin	41,710	5.1	√	-	√	–1.6	1.6	Cytoskeleton/mobility
D7	Keratin type II cytoskeletal 8	53,671	5.3	-	–	√	1.7	–1.7	Cytoskeleton/mobility
D8	Complement componen 1 Q subcomponent-bindin protein, mitochondria	31,342	4.5	-	–	√	3.7	–3.7	Unknown
D9	Tropomyosin alpha-		4.5	-	–	√	–3.1	3.1	Cytoskeleton/mobility
D10	Tropomyosin alpha				–	√	–2.0	2.0	Cytoskeleton/mobility
D11	Alpha-enolase				√	√	1.7	–1.7	Metabolism
D12	Alpha-enolase					√	1.7	–1.7	Metabolism
D13	Hemoglobin beta chain	15,9						–3.8	Binding protein/folding
D14	Non-POU domain-containing octamer-binding protein	54,197	9.4						Binding protein/folding
D15	Unnamed protein product	65,980	7.6	-	–			.2	Unknown
D16	Alpha-enolase	47,008	7.2	-	√	√	1.	1.7	Metabolism
D17	Translationally-controlled tumor protein	19,582	4.6	-	–	√	1.1	–1.1	Binding protein/folding
D18	DNA-binding protein	35,801	8.9	-	–	√	2.0	–2.0	Binding protein/folding
D19	Unnamed protein product	59,492	5.2	-	–	-	–1.7	1.7	Unknown
D20	Heterogeneous nuclear ribonucleoprotein A1	38,822	9.3	-	–	√	2.8	–2.8	Protein synthesis/degradation
D21	Vimentin	53,681	5.0	-	√	√	–6.1	6.1	Cytoskeleton/mobility
D22	Peroxiredoxin 1	22,096	8.2	-	√	√	1.8	–1.8	Protection/detoxification
D23	Peptidyl-prolyl cis-trans isomerase A, cyclophilin	17,981	7.4	-	√	√	1.4	–1.4	Binding protein/folding
D24	78 kDa Glucose-regulated protein	72,377	5.1	-	√	√	–1.0	1.0	Chaperone/stress response
D25	Unnamed protein product	65,980	7.6	-	–	-	1.5	–1.5	Unknown
D26	Cofilin-1	18,490	8.2	-	√	√	–1.9	1.9	Cytoskeleton/mobility

Where does the annotation come from ?
What's about the annotation accuracy ?

Comparative proteomic analysis of lung cancer cell line and lung fibroblast cell line (PMID: 19657000)

Life of sequences …



RNA, genes, genomes, …

Nucleic acid sequence databases

Annotated coding sequences (CDS) / gene prediction

Open access

Protein sequence databases

Open access

Protein sequences + biological knowledge (including GO terms)

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals

- Nucleic acid sequence databases
  - INSDC, RefSeq, Ensembl
- Protein sequence databases
- UniProtKB
  - UniProtKB/Swiss-Prot
    - Protein sequences
    - Biological knowledge
  - UniProtKB/TrEMBL
    - Protein sequences
    - Biological knowledge
  - GO annotation
  - UniProt Proteomes
- NCBI protein (RefSeq)
- Practicals

Protein sequence data: where does it come from ?

- > 190 billion ‘different’ proteins on earth (∑ N species x M genes)

- ~ 130 million ‘known and public’ protein sequences in now
  - 50 % more by next year !

- More than 98% of the protein sequences are derived from the translation of nucleotide sequences (mRNA or DNA/genome)

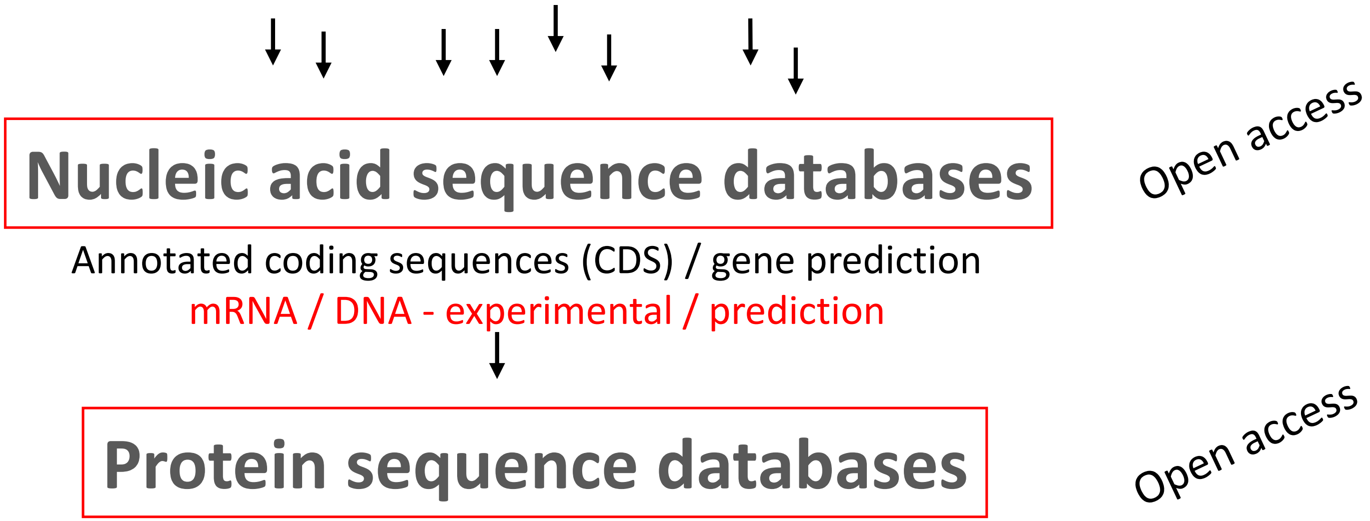
- Less than 1 % come from direct protein sequencing (Edman, MS/MS…)

UniProt

SIB Swiss Institute of Bioinformatics

Life of sequences ...

RNA, genes, genomes, ...



Nucleic acid sequence databases

Open access

Annotated coding sequences (CDS) / gene prediction

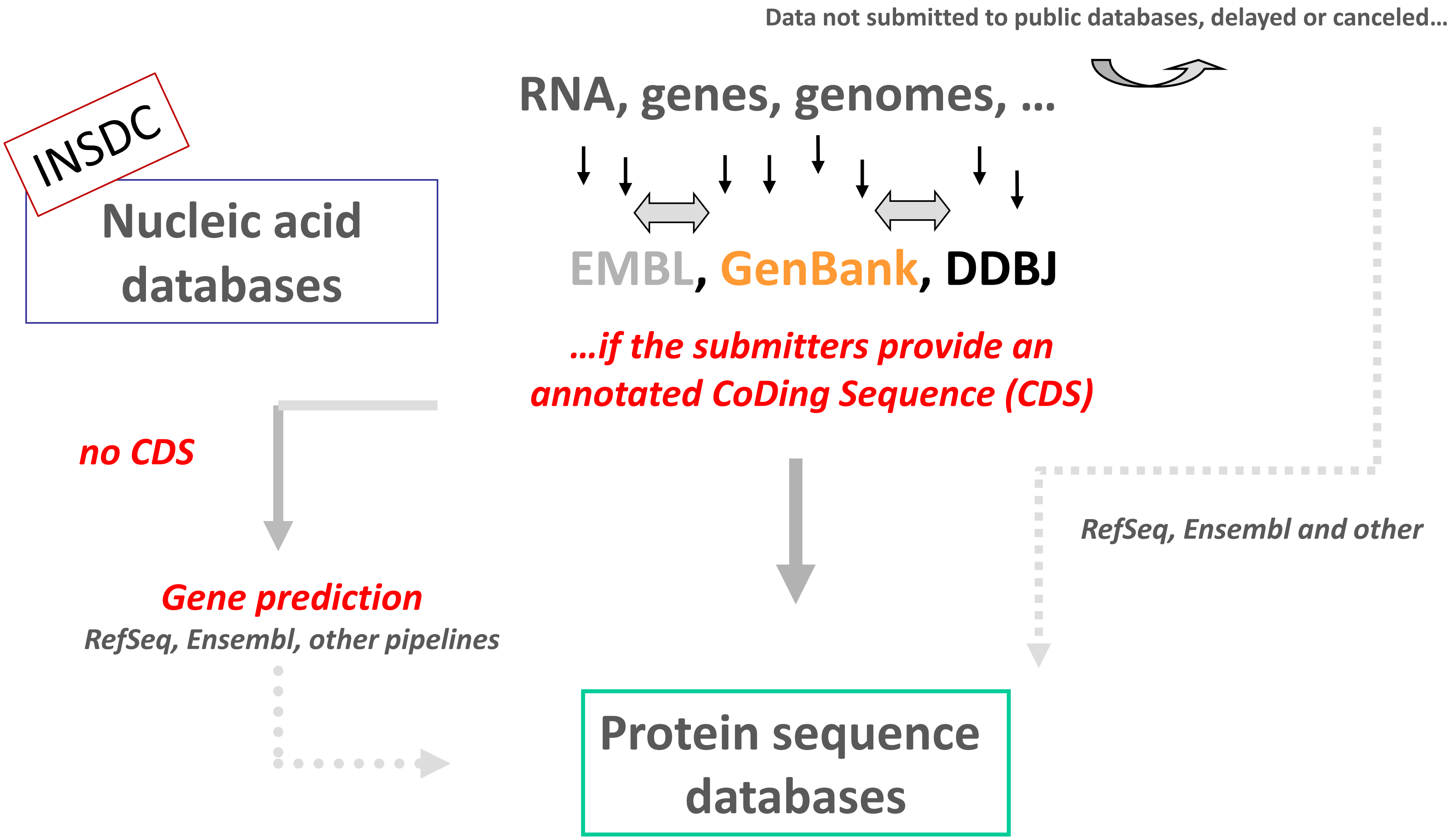
mRNA / DNA - experimental / prediction

Protein sequence databases

Open access

Protein sequences + biological knowledge (including GO terms)

Life of sequences …



Data not submitted to public databases, delayed or canceled…

RNA, genes, genomes, …

INSDC

Nucleic acid databases

EMBL, GenBank, DDBJ

…if the submitters provide an annotated CoDing Sequence (CDS)

no CDS

RefSeq, Ensembl and other

Gene prediction

RefSeq, Ensembl, other pipelines

Protein sequence databases

Nucleic acid sequence databases

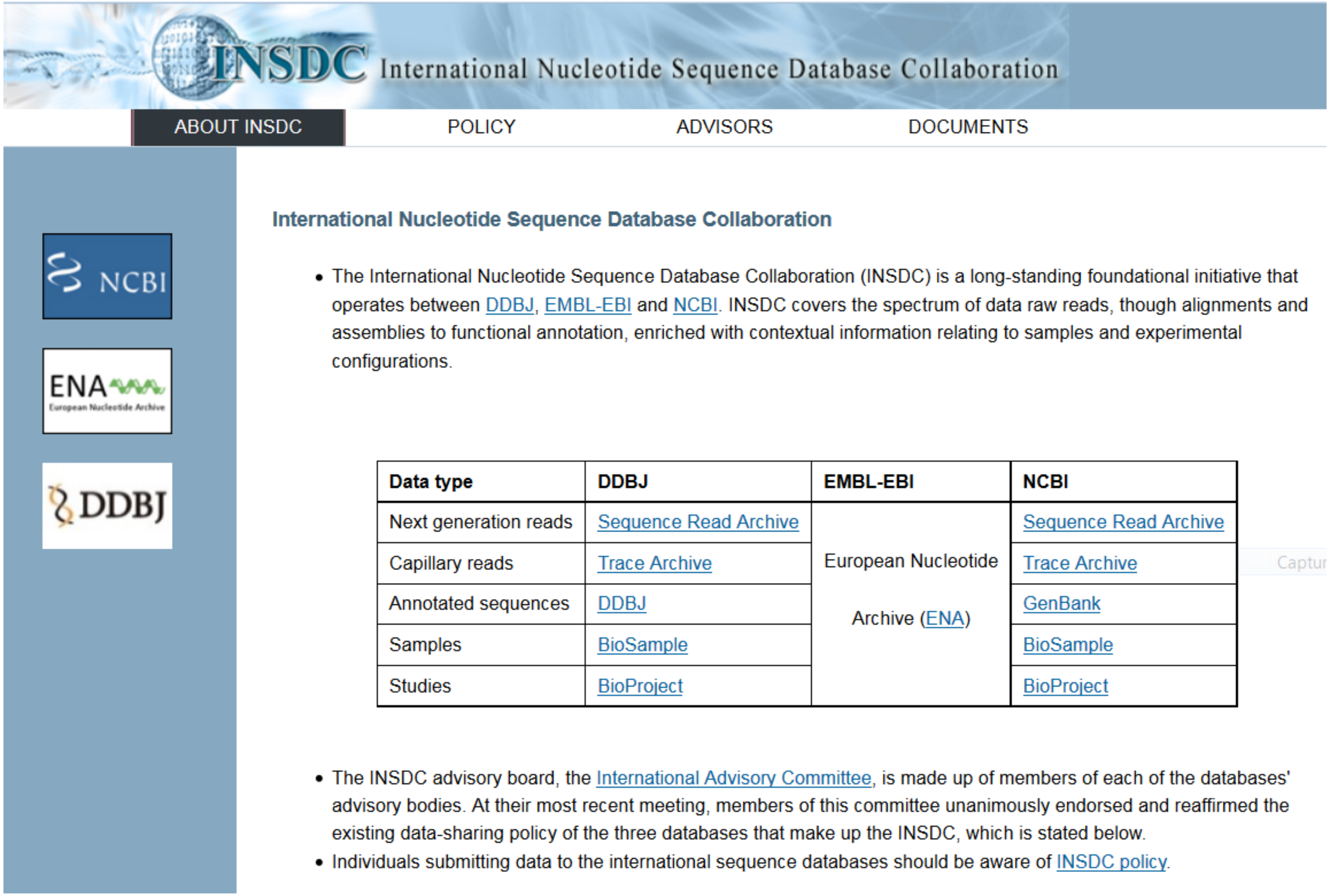
open access

INSDC: ENA, GenBank, DDBJ

RefSeq (NCBI)

Ensembl (EBI)

EMBL-ENA/GenBank/DDBJ



INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
NCBI
ENA European Nucleotide Archive
DDBJ

International Nucleotide Sequence Database Collaboration

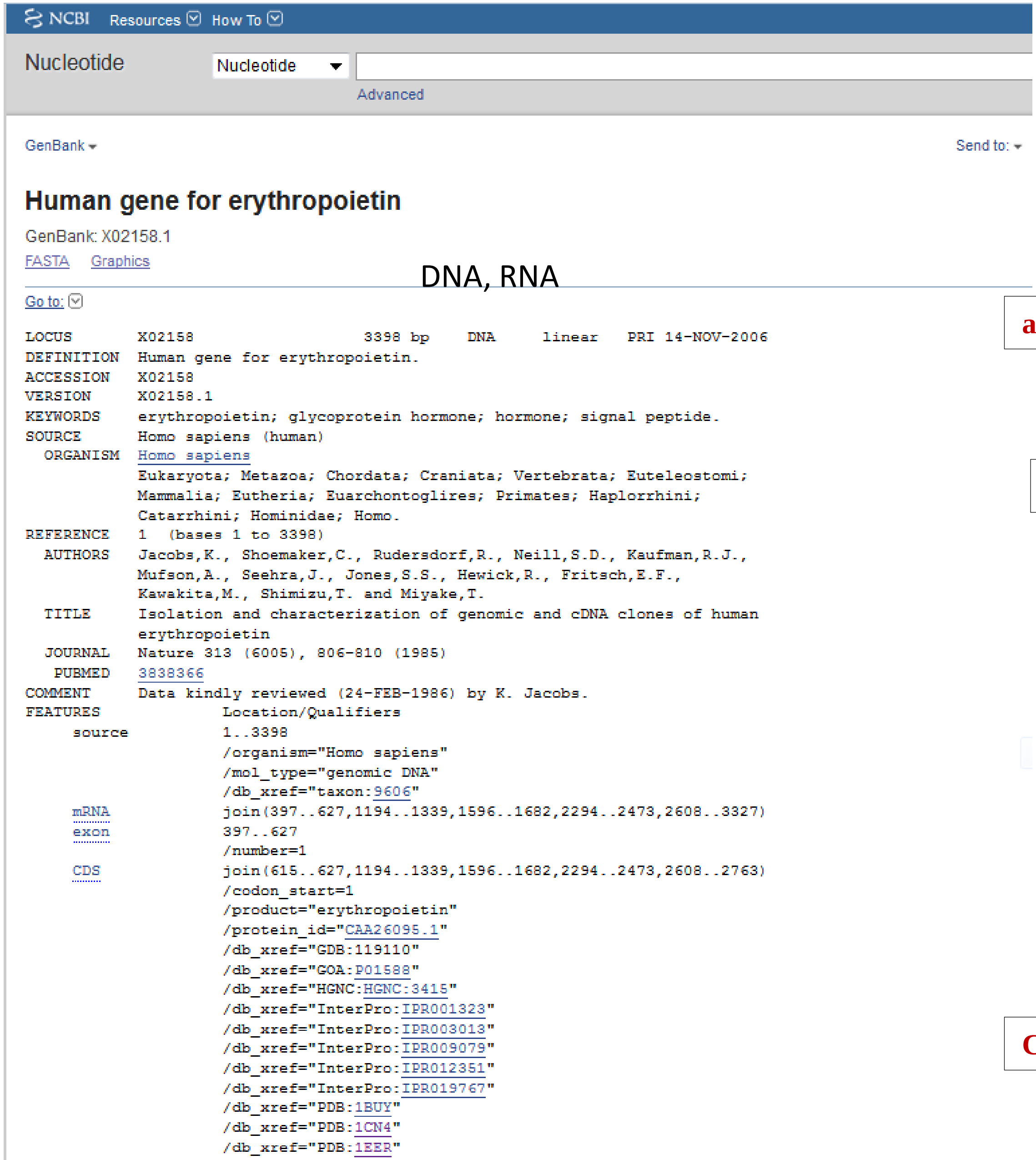
- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI. INSDC covers the spectrum of data raw reads, though alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- The INSDC advisory board, the International Advisory Committee, is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should be aware of INSDC policy.

http://www.insdc.org/

GenBank X02158



Human gene for erythropoietin

GenBank: X02158.1

FASTA Graphics

DNA, RNA

Go to:

```
LOCUS       X02158                  3398 bp    DNA     linear   PRI 14-NOV-2006
DEFINITION  Human gene for erythropoietin.
ACCESSION   X02158
VERSION     X02158.1
KEYWORDS    erythropoietin; glycoprotein hormone; hormone; signal peptide.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3398)
  AUTHORS   Jacobs,K., Shoemaker,C., Rudersdorf,R., Neill,S.D., Kaufman,R.J.,
            Mufson,A., Seehra,J., Jones,S.S., Hewick,R., Fritsch,E.F.,
            Kawakita,M., Shimizu,T. and Miyake,T.
  TITLE     Isolation and characterization of genomic and cDNA clones of human
            erythropoietin
  JOURNAL   Nature 313 (6005), 806-810 (1985)
   PUBMED   3838366
COMMENT     Data kindly reviewed (24-FEB-1986) by K. Jacobs.
FEATURES             Location/Qualifiers
     source          1..3398
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     mRNA            join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)
     exon            397..627
                     /number=1
     CDS             join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
                     /codon_start=1
                     /product="erythropoietin"
                     /protein_id="CAA26095.1"
                     /db_xref="GDB:119110"
                     /db_xref="GOA:P01588"
                     /db_xref="HGNC:HGNC:3415"
                     /db_xref="InterPro:IPR001323"
                     /db_xref="InterPro:IPR003013"
                     /db_xref="InterPro:IPR009079"
                     /db_xref="InterPro:IPR012351"
                     /db_xref="InterPro:IPR019767"
                     /db_xref="PDB:1BUY"
                     /db_xref="PDB:1CN4"
                     /db_xref="PDB:1EER"
```

accession number

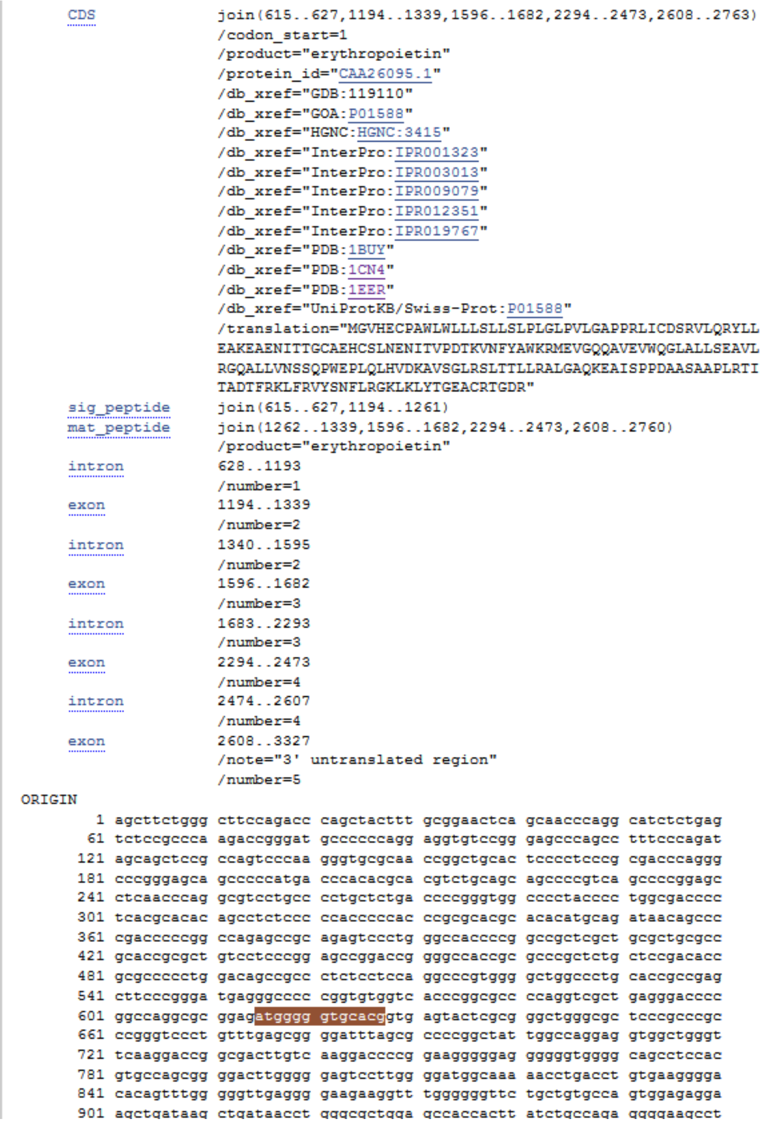
taxonomy

references - the submitters

Cross-references

https://www.ncbi.nlm.nih.gov/nuccore/X02158

GenBank X02158



```
     CDS             join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
                     /codon_start=1
                     /product="erythropoietin"
                     /protein_id="CAA26095.1"
                     /db_xref="GDB:119110"
                     /db_xref="GOA:P01588"
                     /db_xref="HGNC:HGNC:3415"
                     /db_xref="InterPro:IPR001323"
                     /db_xref="InterPro:IPR003013"
                     /db_xref="InterPro:IPR009079"
                     /db_xref="InterPro:IPR012351"
                     /db_xref="InterPro:IPR019767"
                     /db_xref="PDB:1BUY"
                     /db_xref="PDB:1CN4"
                     /db_xref="PDB:1EER"
                     /db_xref="UniProtKB/Swiss-Prot:P01588"
                     /translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
                     EAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVL
                     RGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTI
                     TADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR"
     sig_peptide     join(615..627,1194..1261)
     mat_peptide     join(1262..1339,1596..1682,2294..2473,2608..2760)
                     /product="erythropoietin"
     intron          628..1193
                     /number=1
     exon            1194..1339
                     /number=2
     intron          1340..1595
                     /number=2
     exon            1596..1682
                     /number=3
     intron          1683..2293
                     /number=3
     exon            2294..2473
                     /number=4
     intron          2474..2607
                     /number=4
     exon            2608..3327
                     /note="3' untranslated region"
                     /number=5
ORIGIN
        1 agcttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag
       61 tctccgccca agaccgggat gccccccagg aggtgtccgg gagcccagcc tttcccagat
      121 agcagctccg ccagtcccaa gggtgcgcaa ccggctgcac tcccctcccg cgacccaggg
      181 cccgggagca gcccccatga cccacacgca cgtctgcagc agccccgtca gccccggagc
      241 ctcaacccag gcgtcctgcc cctgctctga ccccgggtgg cccctacccc tggcgacccc
      301 tcacgcacac agcctctccc ccaccccac ccgcgcacgc acacatgcag ataacagccc
      361 cgacccccgg ccagagccgc agagtccctg ggccaccccg gccgctcgct gcgctgcgcc
      421 gcaccgcgct gtcctcccgg agccggaccg gggccaccgc gcccgctctg ctccgacacc
      481 gcgcccctg gacagccgcc ctctcctcca ggcccgtggg gctggccctg caccgccgag
      541 cttcccggga tgagggcccc cggtgtggtc acccggcgcc ccaggtcgct gagggacccc
      601 ggccaggcgc ggagatgggg gtgcacggtg agtactcgcg ggctgggcgc tcccgcccgc
      661 ccgggtccct gtttgagcgg ggatttagcg ccccggctat tggccaggag gtggctgggt
      721 tcaaggaccg gcgacttgtc aaggaccccg gaagggggag gggggtgggg cagcctccac
      781 gtgccagcgg ggacttgggg gagtccttgg ggatggcaaa aacctgacct gtgaagggga
      841 cacagtttgg gggttgaggg gaagaaggtt tggggggttc tgctgtgcca gtggagagga
      901 agctgataag ctgataacct gggcgctgga gccaccactt atctgccaga ggggaagcct
```

CDS annotation
CoDing Sequence
(proposed by submitters)

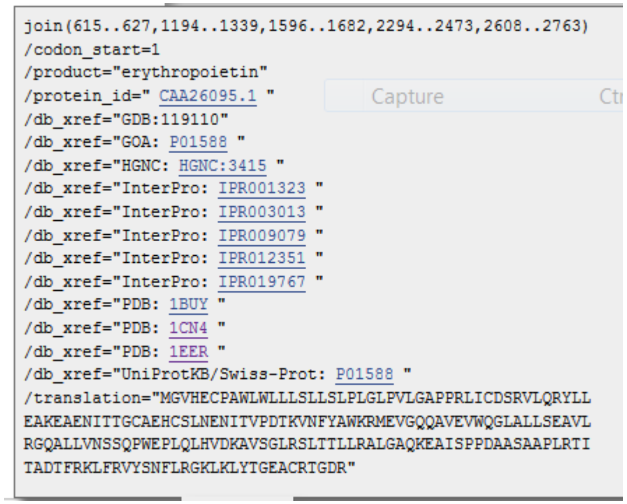
CDS translation

DNA sequence

https://www.ncbi.nlm.nih.gov/nuccore/X02158

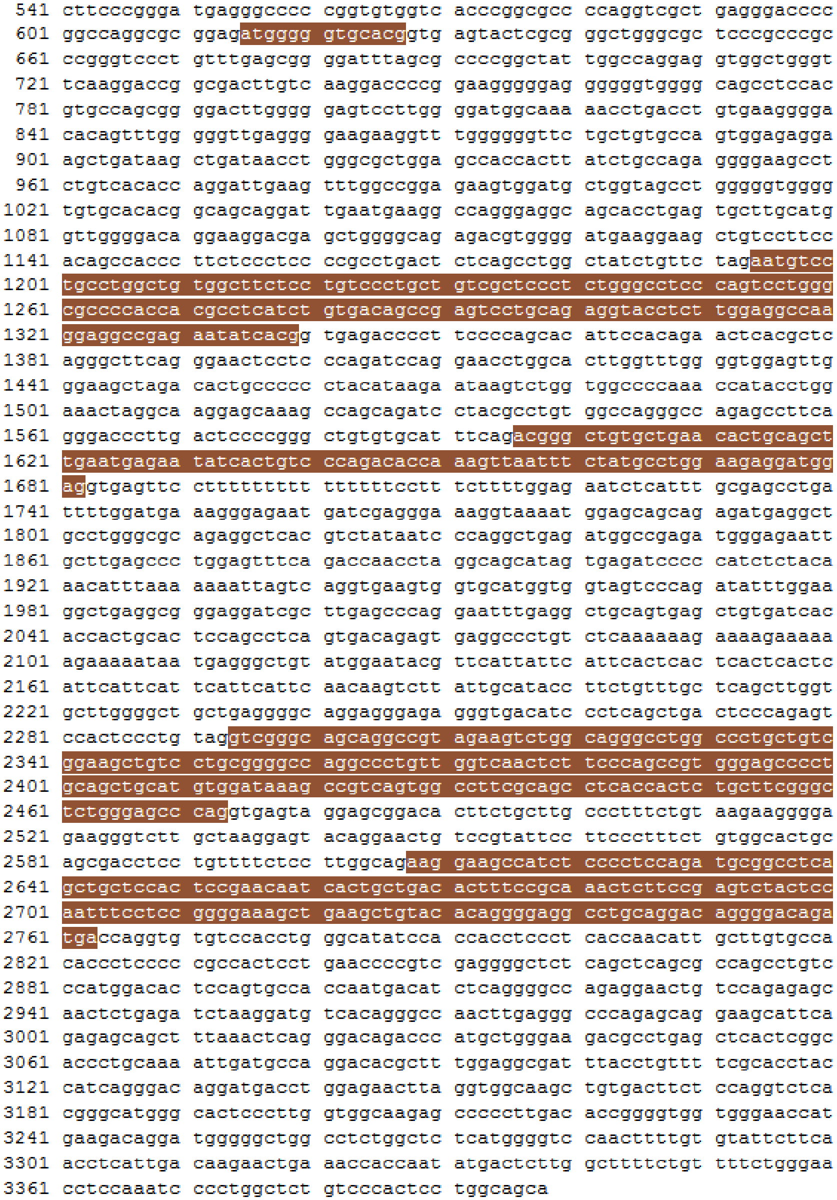
GenBank X02158

CDS
CoDing Sequence
(proposed by submitters)



```
join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
/codon_start=1
/product="erythropoietin"
/protein_id=" CAA26095.1 "
/db_xref="GDB:119110"
/db_xref="GOA: P01588 "
/db_xref="HGNC: HGNC:3415 "
/db_xref="InterPro: IPR001323 "
/db_xref="InterPro: IPR003013 "
/db_xref="InterPro: IPR009079 "
/db_xref="InterPro: IPR012351 "
/db_xref="InterPro: IPR019767 "
/db_xref="PDB: 1BUY "
/db_xref="PDB: 1CN4 "
/db_xref="PDB: 1EER "
/db_xref="UniProtKB/Swiss-Prot: P01588 "
/translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
EAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVL
RGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTI
TADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR"
```

either generated by gene prediction programs or experimentally proven.



```
 541 cttcccggga tgaggcccc cggtgtggtc acccggcgcc ccaggtcgct gagggacccc
 601 ggccaggcgc ggagatgggg gtgcacggtg agtactcgcg ggctgggcgc tcccgcccgc
 661 ccgggtccct gtttgagcgg ggatttagcg ccccggctat tggccaggag gtggctgggt
 721 tcaaggaccg gcgacttgtc aaggaccccg gaagggggag gggggtgggg cagcctccac
 781 gtgccagcgg ggacttgggg gagtccttgg ggatggcaaa aacctgacct gtgaagggga
 841 cacagtttgg gggttgaggg gaagaaggtt tgggggttc tgctgtgcca gtggagagga
 901 agctgataag ctgataacct gggcgctgga gccaccactt atctgccaga ggggaagcct
 961 ctgtcacacc aggattgaag tttggccgga gaagtggatg ctggtagcct gggggtgggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccagggaggc agcacctgag tgcttgcatg
1081 gttggggaca ggaaggacga gctggggcag agacgtgggg atgaaggaag ctgtccttcc
1141 acagccaccc ttctccctcc ccgcctgact ctcagcctgg ctatctgttc tagaatgtcc
1201 tgcctggctg tggcttctcc tgtccctgct gtcgctccct ctgggcctcc cagtcctggg
1261 cgccccacca cgcctcatct gtgacagccg agtcctgcag aggtacctct tggaggccaa
1321 ggaggccgag aatatcacgg tgagacccct tccccagcac attccacaga actcacgctc
1381 agggcttcag ggaactcctc ccagatccag gaacctggca cttggtttgg ggtggagttg
1441 ggaagctaga cactgccccc ctacataaga ataagtctgg tggccccaaa ccatacctgg
1501 aaactaggca aggagcaaag ccagcagatc ctacgcctgt ggccagggcc agagccttca
1561 gggacccttg actccccggg ctgtgtgcat ttcagacggg ctgtgctgaa cactgcagct
1621 tgaatgagaa tatcactgtc ccagacacca aagttaattt ctatgcctgg aagaggatgg
1681 aggtgagttc cttttttttt ttttttcctt tcttttggag aatctcattt gcgagcctga
1741 ttttggatga aagggagaat gatcgaggga aaggtaaaat ggagcagcag agatgaggct
1801 gcctgggcgc agaggctcac gtctataatc ccaggctgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtttca gaccaaccta ggcagcatag tgagatcccc catctctaca
1921 aacatttaaa aaaattagtc aggtgaagtg gtgcatggtg gtagtcccag atatttggaa
1981 ggctgaggcg ggaggatcgc ttgagcccag gaatttgagg ctgcagtgag ctgtgatcac
2041 accactgcac tccagcctca gtgacagagt gaggccctgt ctcaaaaaag aaaagaaaaa
2101 agaaaaataa tgagggctgt atggaatacg ttcattattc attcactcac tcactcactc
2161 attcattcat tcattcattc aacaagtctt attgcatacc ttctgtttgc tcagcttggt
2221 gcttggggct gctgaggggc aggagggaga gggtgacatc cctcagctga ctcccagagt
2281 ccactccctg taggtcgggc agcaggccgt agaagtctgg cagggcctgg ccctgctgtc
2341 ggaagctgtc ctgcggggcc aggccctgtt ggtcaactct tcccagccgt gggagcccct
2401 gcagctgcat gtggataaag ccgtcagtgg ccttcgcagc ctcaccactc tgcttcgggc
2461 tctgggagcc caggtgagta ggagcggaca cttctgcttg ccctttctgt aagaagggga
2521 gaagggtctt gctaaggagt acaggaactg tccgtattcc ttccctttct gtggcactgc
2581 agcgacctcc tgttttctcc ttggcagaag gaagccatct cccctccaga tgcggcctca
2641 gctgctccac tccgaacaat cactgctgac actttccgca aactcttccg agtctactcc
2701 aatttcctcc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaccaggtg tgtccacctg ggcatatcca ccacctccct caccaacatt gcttgtgcca
2821 caccctcccc cgccactcct gaaccccgtc gaggggctct cagctcagcg ccagcctgtc
2881 ccatggacac tccagtgcca ccaatgacat ctcaggggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tcacagggcc aacttgaggg cccagagcag gaagcattca
3001 gagagcagct ttaaactcag ggacagaccc atgctgggaa gacgcctgag ctcactcggc
3061 accctgcaaa attgatgcca ggacacgctt tggaggcgat ttacctgttt tcgcacctac
3121 catcagggac aggatgacct ggagaactta ggtggcaagc tgtgacttct ccaggtctca
3181 cgggcatggg cactcccttg gtggcaagag ccccccttgac accggggtgg tgggaaccat
3241 gaagacagga tgggggctgg cctctggctc tcatggggtc caacttttgt gtattcttca
3301 acctcattga caagaactga aaccaccaat atgactcttg gcttttctgt tttctgggaa
3361 cctccaaatc ccctggctct gtcccactcc tggcagca
```

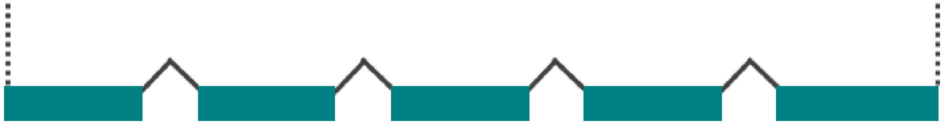
https://www.ncbi.nlm.nih.gov/nuccore/x02158

Coding sequence (CDS) annotation

submit

DNA sequence

ACGCTCGTACGCATCGTCACTACTAGCTACGACGACGACACGCTACTACTCGACGATTCT



CDS
CoDing Sequence
(provided by submitters)

transcribe

Derived mRNA sequence

AUGCGUAGUGAUGAAUGCUGCUGUGCGAUGAGCUGC

translate

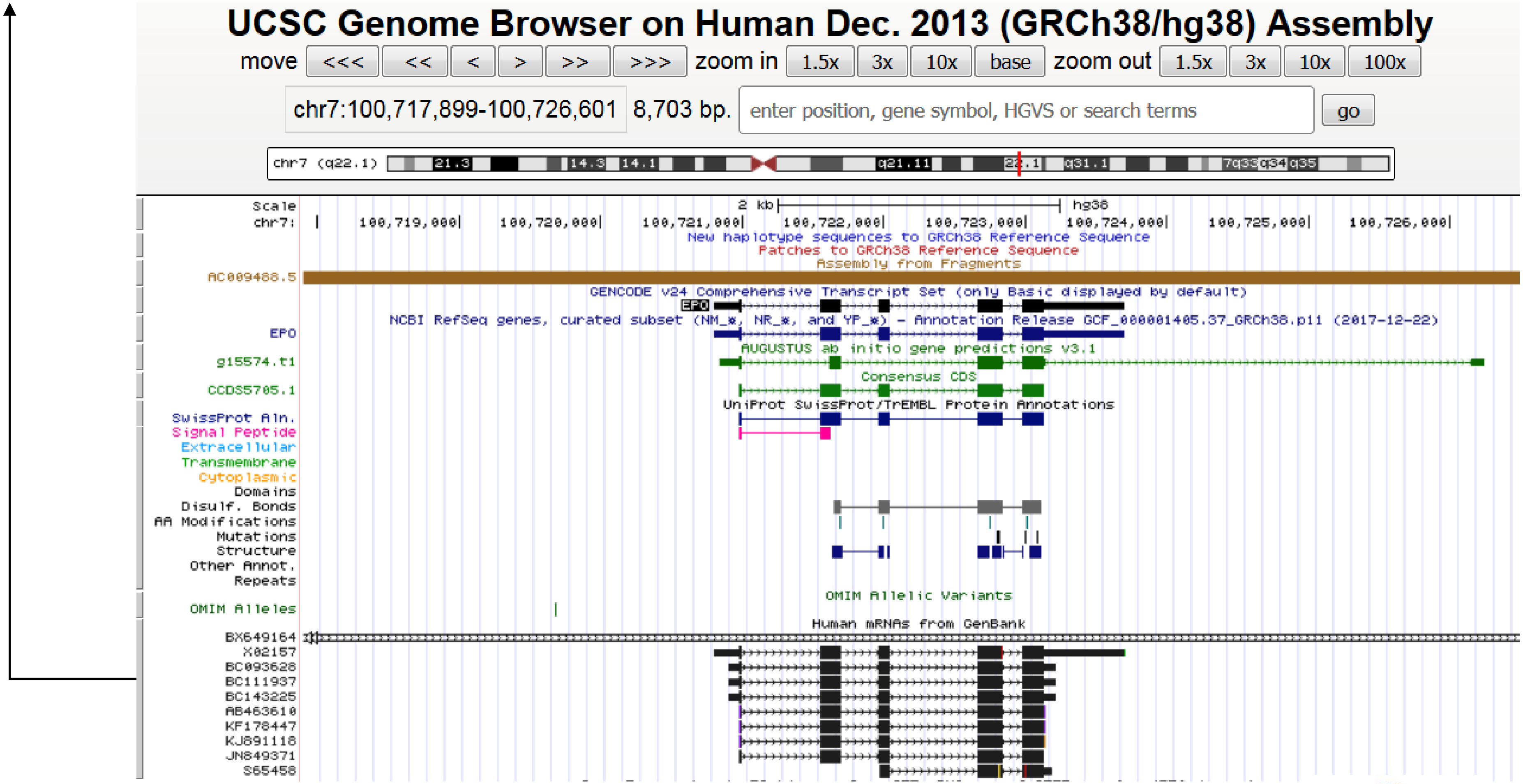
Derived protein sequence

MRSNECCCAMSC

Slide J. McDowall

UCSC genome browser: human EPO

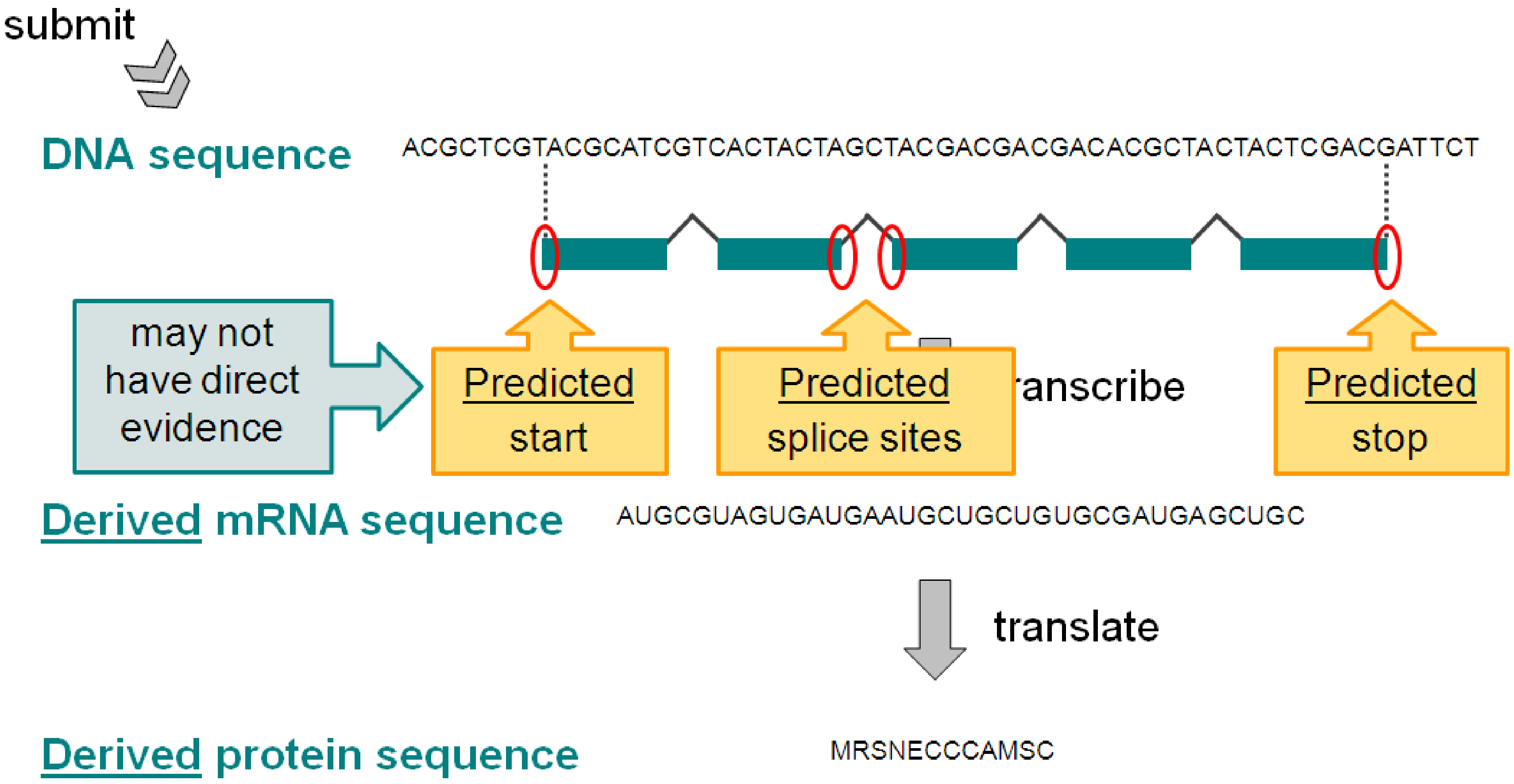
mRNAs and their corresponding CDS annotation
(from EMBL/GenBank/DDBJ)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

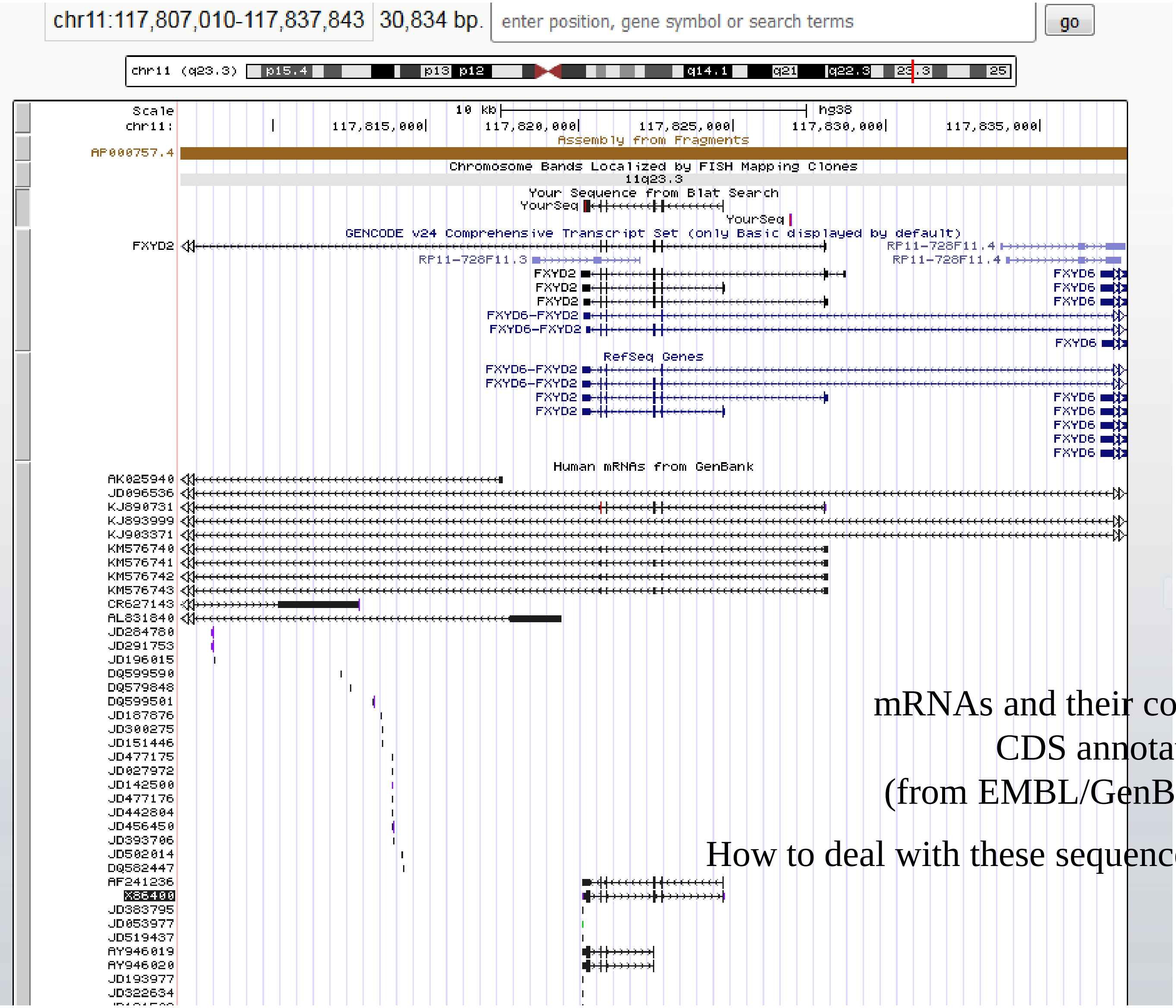
chr7:100,717,899-100,726,601 8,703 bp.

Coding sequence (CDS) annotation



Slide J. McDowall

UCSC genome browser: another gene…



chr11:117,807,010-117,837,843 30,834 bp.

mRNAs and their corresponding CDS annotation (from EMBL/GenBank/DDBJ)

How to deal with these sequences: see later….

EMBL/GenBank/DDBJ

- Archive: nothing goes out -> highly redundant !

- Most annotations (including CDS, gene and protein names, …) are done by the submitters: heterogeneity of the quality and of the completion

- Many errors: in sequences, in annotations (gene and protein names, …), in CDS attribution, no consistency of annotations

- Archive: all submitted information remains there; not updated

EMBL/GenBank/DDBJ and annotation

“Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), **the quality and accuracy of the record are the responsibility of the submitting author, not of the database.** The databases will work with submitters and users of the database to achieve the best quality resource possible.”

http://www.insdc.org/policy

EMBL/GenBank/DDBJ and annotation

- many scientists assume that GenBank annotation is kept up to date, and they are surprised to hear that it is not
- the **annotation has remained static**: a gene labeled 'hypothetical protein' a few years ago might now have a known function.
- **erroneous and inconsistent naming of genes**.
- a name is **transferred from one gene to another** on the basis of sequence similarity (usually from a BLAST search). As more genomes are annotated, and more BLAST searches are run, **the original source of the name quickly becomes lost**.
- scientists should fix errors that they find. But this would quickly destroy the **archival function of GenBank**, as original entries would be erased over time.

(PMID: 17274839)

EMBL/GenBank/DDBJ and UniProtKB

What is transferred to UniProtKB:

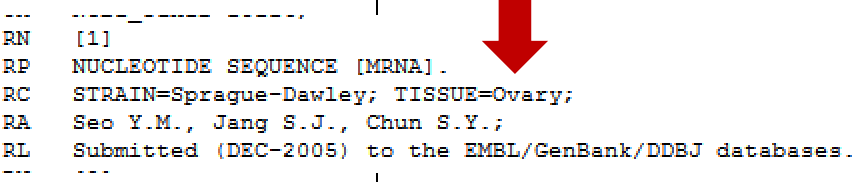
- The protein sequence (translated CDS)

- Gene and protein names
- Origin of the sequence (tissues)
- Taxonomy
- Publication provided by the submitting author

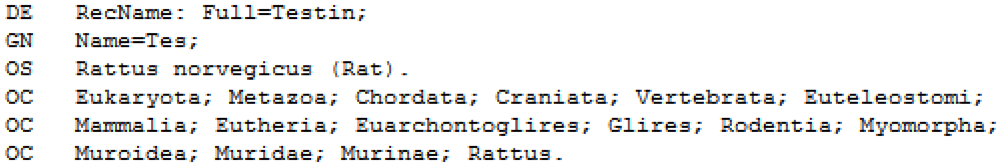
What is transferred to UniProtKB (‘imported’):

DR EMBL; DQ339047; ABC68418.1; -; mRNA.

FT source 1..1397
FT /organism="Rattus norvegicus"
FT /strain="Sprague-Dawley"
FT /mol_type="mRNA"
FT /sex="female"
FT /tissue_type="ovary"
FT /db_xref="taxon:10116"
FT CDS 70..1329
FT /codon_start=1
FT /product="testis derived transcript"
FT /note="TES"
FT /db_xref="GOA:Q2LAP6"



```
---
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   STRAIN=Sprague-Dawley; TISSUE=Ovary;
RA   Seo Y.M., Jang S.J., Chun S.Y.;
RL   Submitted (DEC-2005) to the EMBL/GenBank/DDBJ databases.
---  ---
```



```
DE   RecName: Full=Testin;
GN   Name=Tes;
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
OC   Muroidea; Muridae; Murinae; Rattus.
```

Nucleic acid sequence databases

open access

INSDC: ENA, GenBank, DDBJ

RefSeq (NCBI)

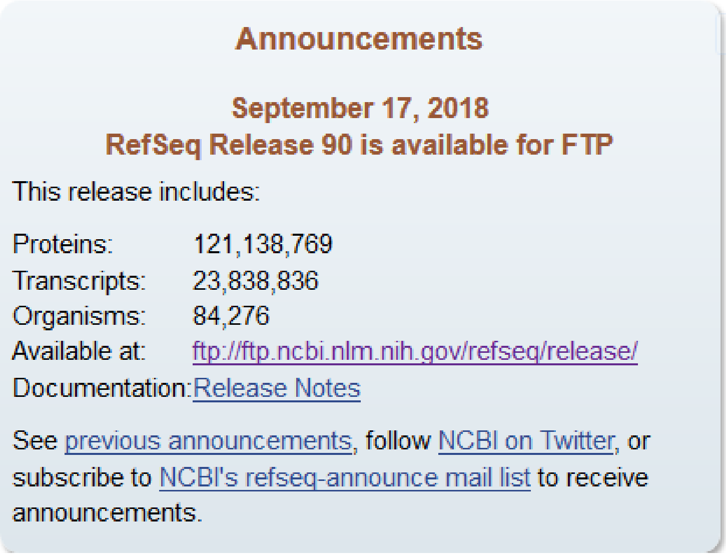
Ensembl (EBI)

RefSeq @NCBI

- The Reference Sequence (RefSeq) database provides a non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

- Contains sequences constructed from INSDC sequences.
- Contains protein sequences derived from gene prediction, not submitted to EMBL/GenBank/DDBJ

One record for one transcript



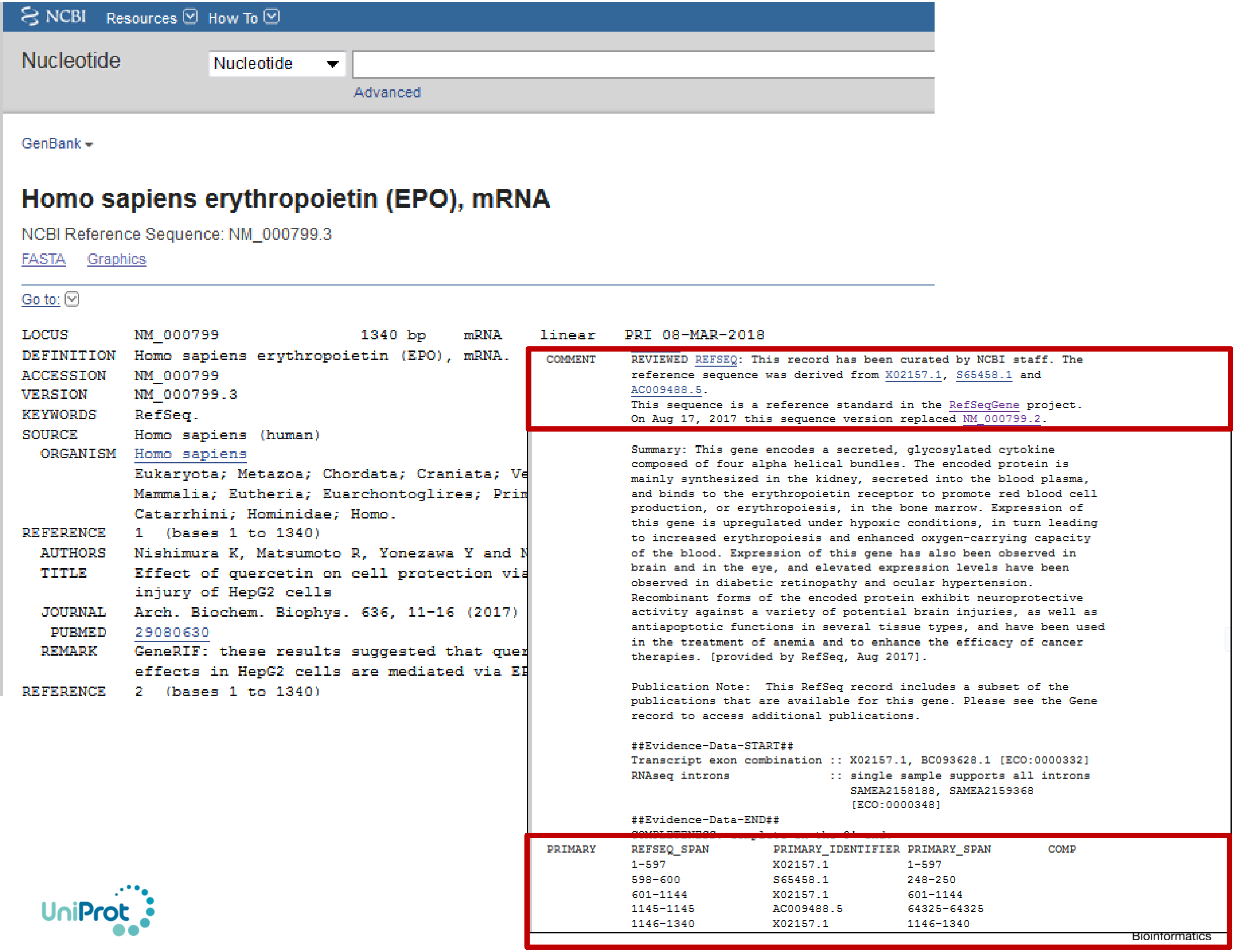
Announcements

September 17, 2018
RefSeq Release 90 is available for FTP

This release includes:

Proteins:	121,138,769
Transcripts:	23,838,836
Organisms:	84,276
Available at:	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/
Documentation:	Release Notes

See previous announcements, follow NCBI on Twitter, or subscribe to NCBI's refseq-announce mail list to receive announcements.



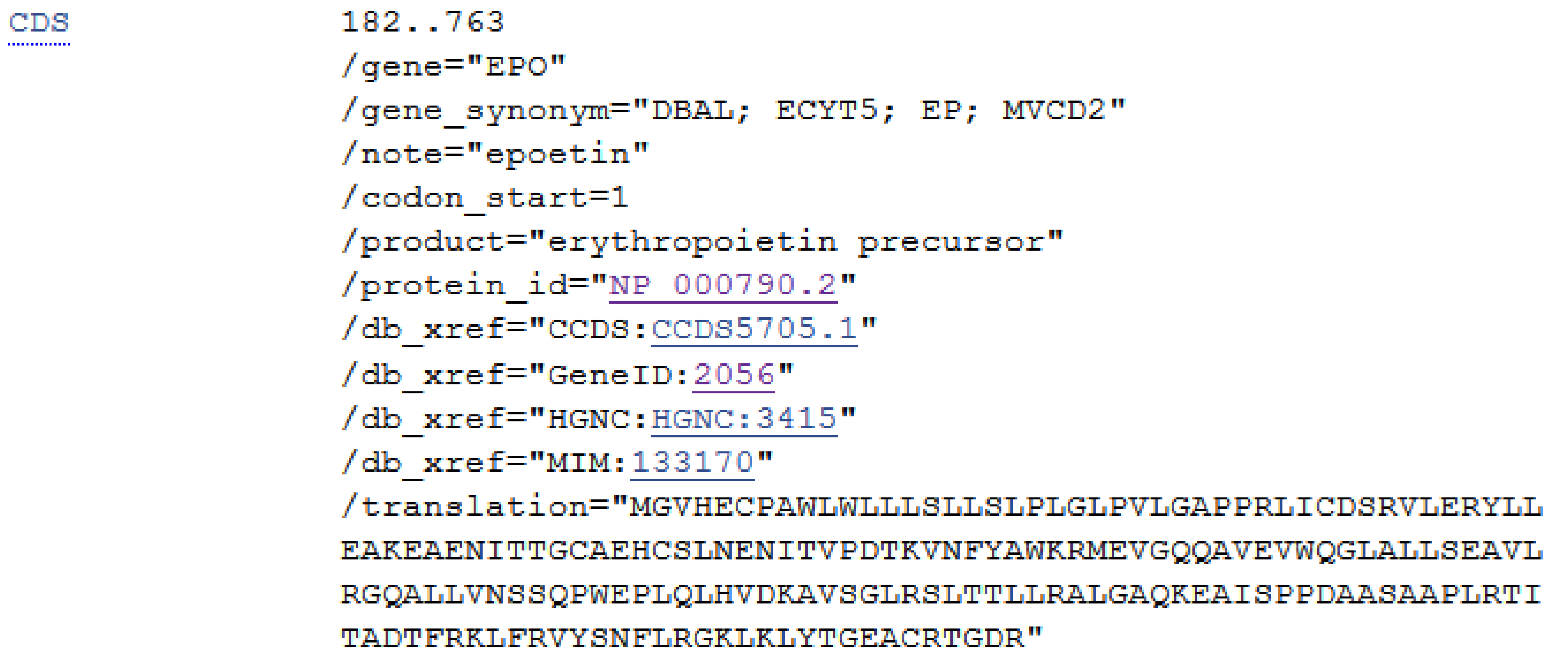
Homo sapiens erythropoietin (EPO), mRNA

NCBI Reference Sequence: NM_000799.3

```
LOCUS       NM_000799               1340 bp    mRNA    linear   PRI 08-MAR-2018
DEFINITION  Homo sapiens erythropoietin (EPO), mRNA.
ACCESSION   NM_000799
VERSION     NM_000799.3
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Ve
            Mammalia; Eutheria; Euarchontoglires; Prim
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1340)
  AUTHORS   Nishimura K, Matsumoto R, Yonezawa Y and N
  TITLE     Effect of quercetin on cell protection via
            injury of HepG2 cells
  JOURNAL   Arch. Biochem. Biophys. 636, 11-16 (2017)
   PUBMED   29080630
  REMARK    GeneRIF: these results suggested that quer
            effects in HepG2 cells are mediated via EP
REFERENCE   2  (bases 1 to 1340)
```

```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from X02157.1, S65458.1 and
            AC009488.5.
            This sequence is a reference standard in the RefSeqGene project.
            On Aug 17, 2017 this sequence version replaced NM_000799.2.

            Summary: This gene encodes a secreted, glycosylated cytokine
            composed of four alpha helical bundles. The encoded protein is
            mainly synthesized in the kidney, secreted into the blood plasma,
            and binds to the erythropoietin receptor to promote red blood cell
            production, or erythropoiesis, in the bone marrow. Expression of
            this gene is upregulated under hypoxic conditions, in turn leading
            to increased erythropoiesis and enhanced oxygen-carrying capacity
            of the blood. Expression of this gene has also been observed in
            brain and in the eye, and elevated expression levels have been
            observed in diabetic retinopathy and ocular hypertension.
            Recombinant forms of the encoded protein exhibit neuroprotective
            activity against a variety of potential brain injuries, as well as
            antiapoptotic functions in several tissue types, and have been used
            in the treatment of anemia and to enhance the efficacy of cancer
            therapies. [provided by RefSeq, Aug 2017].

            Publication Note:  This RefSeq record includes a subset of the
            publications that are available for this gene. Please see the Gene
            record to access additional publications.

            ##Evidence-Data-START##
            Transcript exon combination :: X02157.1, BC093628.1 [ECO:0000332]
            RNAseq introns              :: single sample supports all introns
                                           SAMEA2158188, SAMEA2159368
                                           [ECO:0000348]
            ##Evidence-Data-END##
            COMPLETENESS: complete on the 3' end.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
            1-597               X02157.1           1-597
            598-600             S65458.1           248-250
            601-1144            X02157.1           601-1144
            1145-1145           AC009488.5         64325-64325
            1146-1340           X02157.1           1146-1340
```

RefSeq: NP_000790.2. NM_000799.2.



```
CDS                 182..763
                    /gene="EPO"
                    /gene_synonym="DBAL; ECYT5; EP; MVCD2"
                    /note="epoetin"
                    /codon_start=1
                    /product="erythropoietin precursor"
                    /protein_id="NP_000790.2"
                    /db_xref="CCDS:CCDS5705.1"
                    /db_xref="GeneID:2056"
                    /db_xref="HGNC:HGNC:3415"
                    /db_xref="MIM:133170"
                    /translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLL
                    EAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVL
                    RGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTI
                    TADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR"
```

One record for one transcript

Nucleic acid sequence databases

open access

INSDC: ENA, GenBank, DDBJ
RefSeq (NCBI)
Ensembl (EBI)

UniProt

SIB Swiss Institute of Bioinformatics

Ensembl

- Creates, integrates and distributes reference datasets

- Joint project between EMBL-EBI and the Sanger Centre

Ensembl ENST00000252723; ENSP00000252723; ENSG00000130427.

- Contains protein sequences derived from gene prediction which are not submitted to EMBL/GenBank/DDBJ

  -> included in UniProtKB

Ensembl and UniProtKB

- UniProtKB contain protein sequences derived from Ensembl gene prediction

https://www.ensembl.org/info/genome/genebuild/automatic_coding.html

- Proteome sets in UniProtKB are made of records with a cross-reference to Ensembl

- See later

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

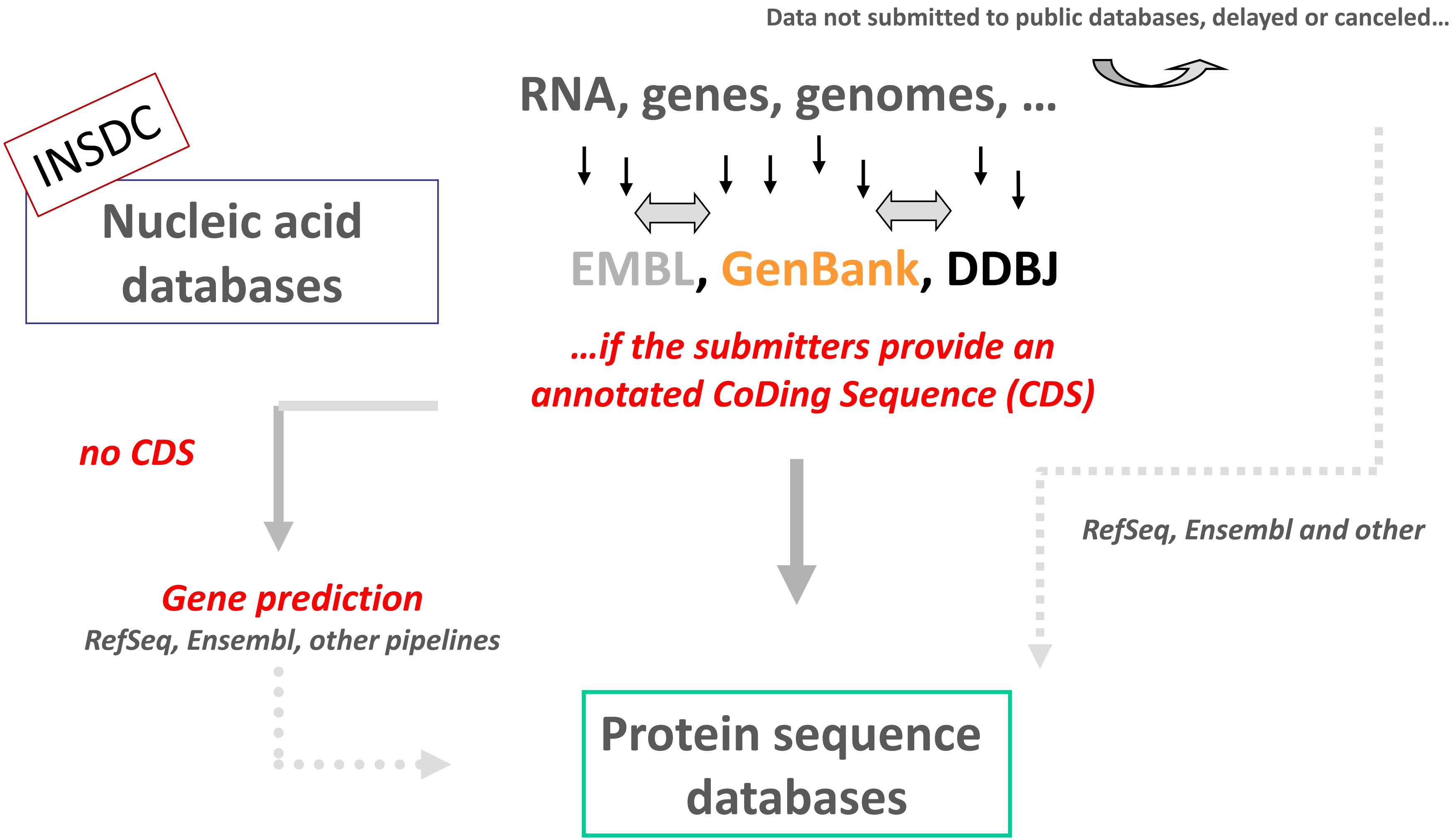
UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals

The life of a sequence …



Data not submitted to public databases, delayed or canceled…

RNA, genes, genomes, …

INSDC

Nucleic acid databases

EMBL, GenBank, DDBJ

…if the submitters provide an annotated CoDing Sequence (CDS)

no CDS

Gene prediction

RefSeq, Ensembl, other pipelines

RefSeq, Ensembl and other

Protein sequence databases

UniProt

SIB Swiss Institute of Bioinformatics

Protein sequence databases

UniProtKB/Swiss-Prot: manually annotated protein sequences (12’500 species)

UniProtKB/TrEMBL: submitted CDS (EMBL-ENA) + automated annotation; non redundant with Swiss-Prot (700’000 species)

GenPept: submitted CDS (GenBank); no annotation; redundant with Swiss-Prot (700’000 species);

PIR: Protein Information Ressource; archive since 2003; integrated into UniProtKB; still available in NCBInr

PDB: Protein Databank: 3D data and associated sequences

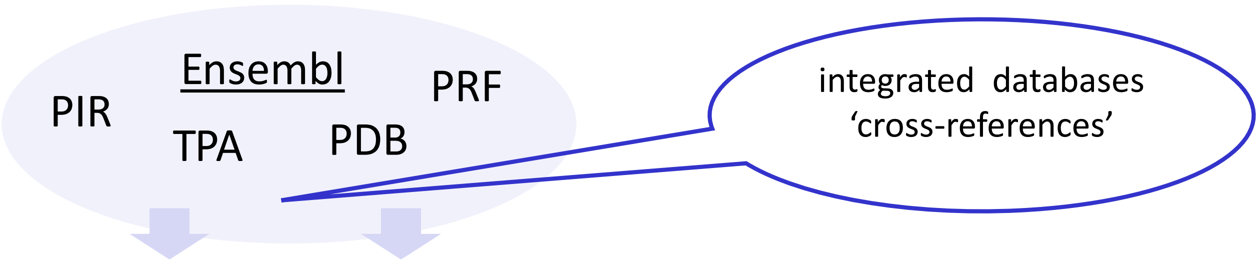
PRF: Journal scan of ‘published’ peptide sequences

RefSeq: Reference Sequence for DNA, RNA, protein + gene prediction + partial manual annotation (68’200 species)

NextProt: human proteins from UniProtKB/Swiss-Prot

NCBInr: Swiss-Prot + GenPept + PIR + PDB + PRF + RefSeq

Major ‘general’ protein sequence database ‘sources’



integrated databases ‘cross-references’

UniProtKB: **Swiss-Prot + TrEMBL**

databases are kept separated

NCBI-nr: **Swiss-Prot + TrEMBL** + **GenPept** + PIR + PDB + PRF + **RefSeq** + TPA

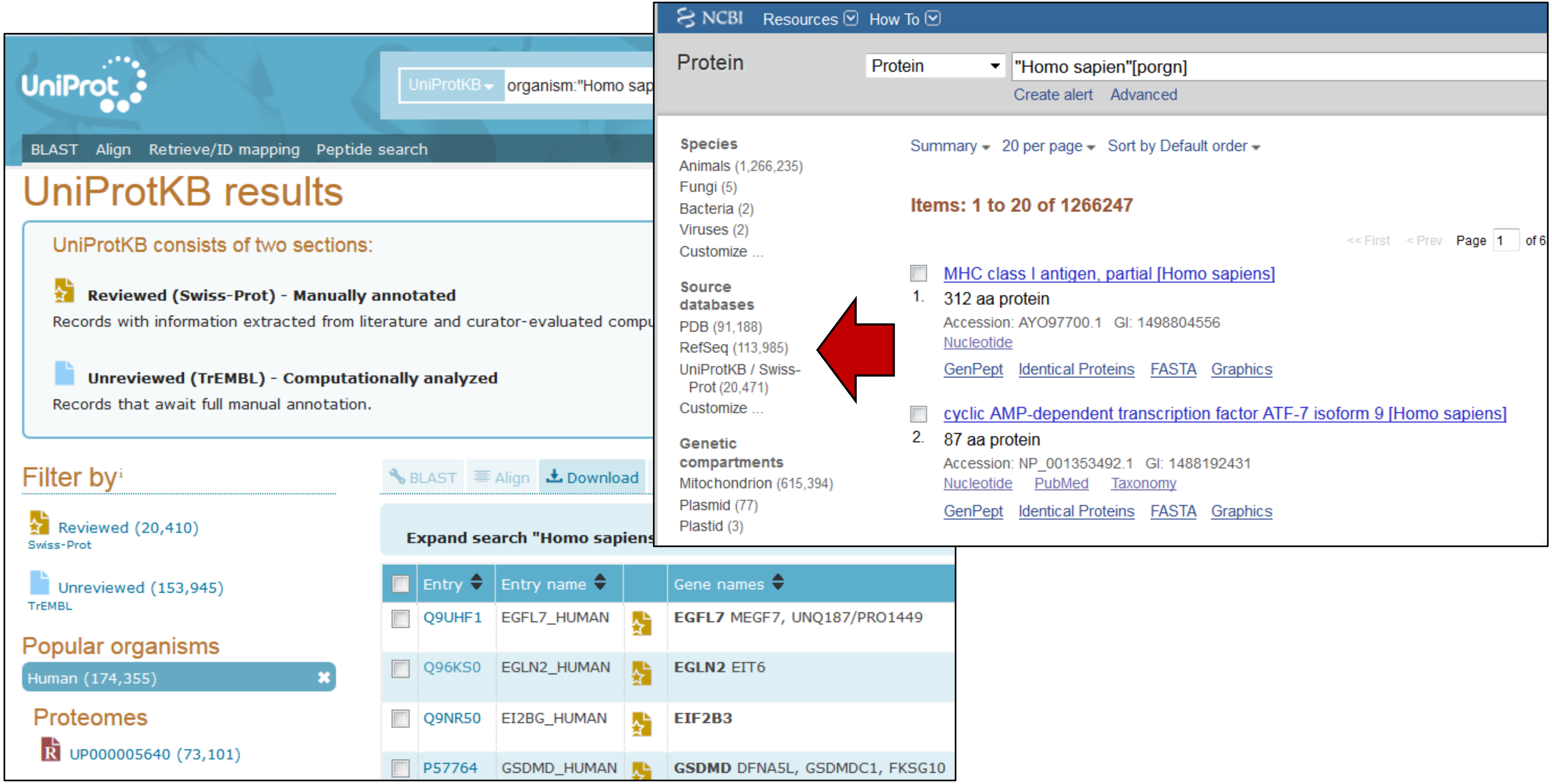
not complete !!!
(only entries created before 2007 ?)

Ensembl and Refseq: gene prediction

Homo sapiens

@ UniProt

@ NCBInr



UniProt

UniProtKB ▾ organism:"Homo sap

BLAST Align Retrieve/ID mapping Peptide search

UniProtKB results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated compu

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

Filter by

Reviewed (20,410)
Swiss-Prot

Unreviewed (153,945)
TrEMBL

Popular organisms

Human (174,355) ✖

Proteomes

UP000005640 (73,101)

BLAST Align Download

Expand search "Homo sapiens

□	Entry	Entry name		Gene names
□	Q9UHF1	EGFL7_HUMAN	★	**EGFL7** MEGF7, UNQ187/PRO1449
□	Q96KS0	EGLN2_HUMAN	★	**EGLN2** EIT6
□	Q9NR50	EI2BG_HUMAN	★	**EIF2B3**
□	P57764	GSDMD_HUMAN	★	**GSDMD** DFNA5L, GSDMDC1, FKSG10

NCBI Resources How To

Protein

Protein ▾ "Homo sapien"[porgn]

Create alert Advanced

Species
Animals (1,266,235)
Fungi (5)
Bacteria (2)
Viruses (2)
Customize ...

Source databases
PDB (91,188)
RefSeq (113,985)
UniProtKB / Swiss-Prot (20,471)
Customize ...

Genetic compartments
Mitochondrion (615,394)
Plasmid (77)
Plastid (3)

Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 1266247

<< First < Prev Page 1 of 6

□ MHC class I antigen, partial [Homo sapiens]
1. 312 aa protein
Accession: AYO97700.1 GI: 1498804556
Nucleotide
GenPept Identical Proteins FASTA Graphics

□ cyclic AMP-dependent transcription factor ATF-7 isoform 9 [Homo sapiens]
2. 87 aa protein
Accession: NP_001353492.1 GI: 1488192431
Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals

www.uniprot.org

~130 millions of proteins/entries

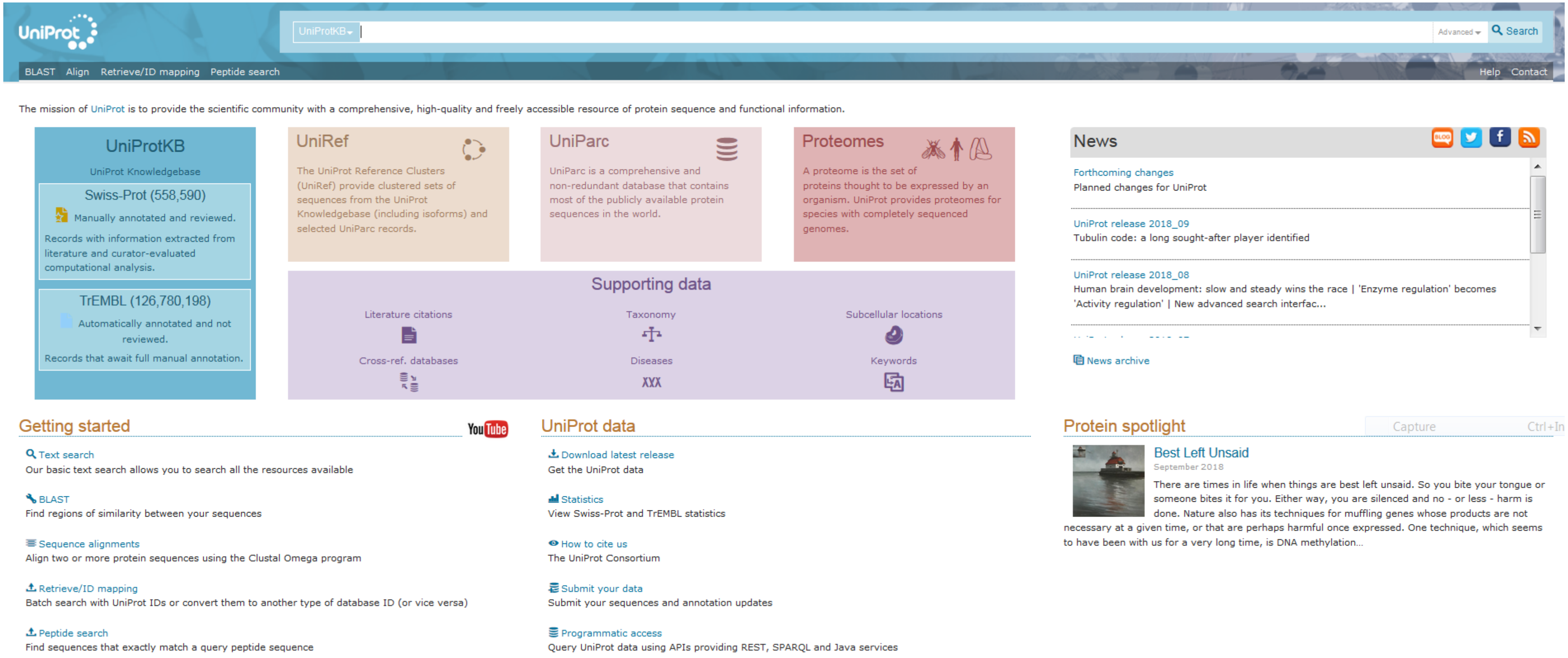
derived from ~ 710’000 different species

7 million unique visitors/year

New release every month

UniProt consortium : EMBL-EBI PIR SIB

EBI : European Bioinformatics Institute (UK)
SIB : Swiss Institute of Bioinformatics (CH)
PIR : Protein Information Resource (USA)



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (558,590)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (126,780,198)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2018_09
Tubulin code: a long sought-after player identified

UniProt release 2018_08
Human brain development: slow and steady wins the race | 'Enzyme regulation' becomes 'Activity regulation' | New advanced search interfac...

News archive

Getting started

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping
Batch search with UniProt IDs or convert them to another type of database ID (or vice versa)

Peptide search
Find sequences that exactly match a query peptide sequence

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotation updates

Programmatic access
Query UniProt data using APIs providing REST, SPARQL and Java services

Protein spotlight

Capture Ctrl+In

Best Left Unsaid

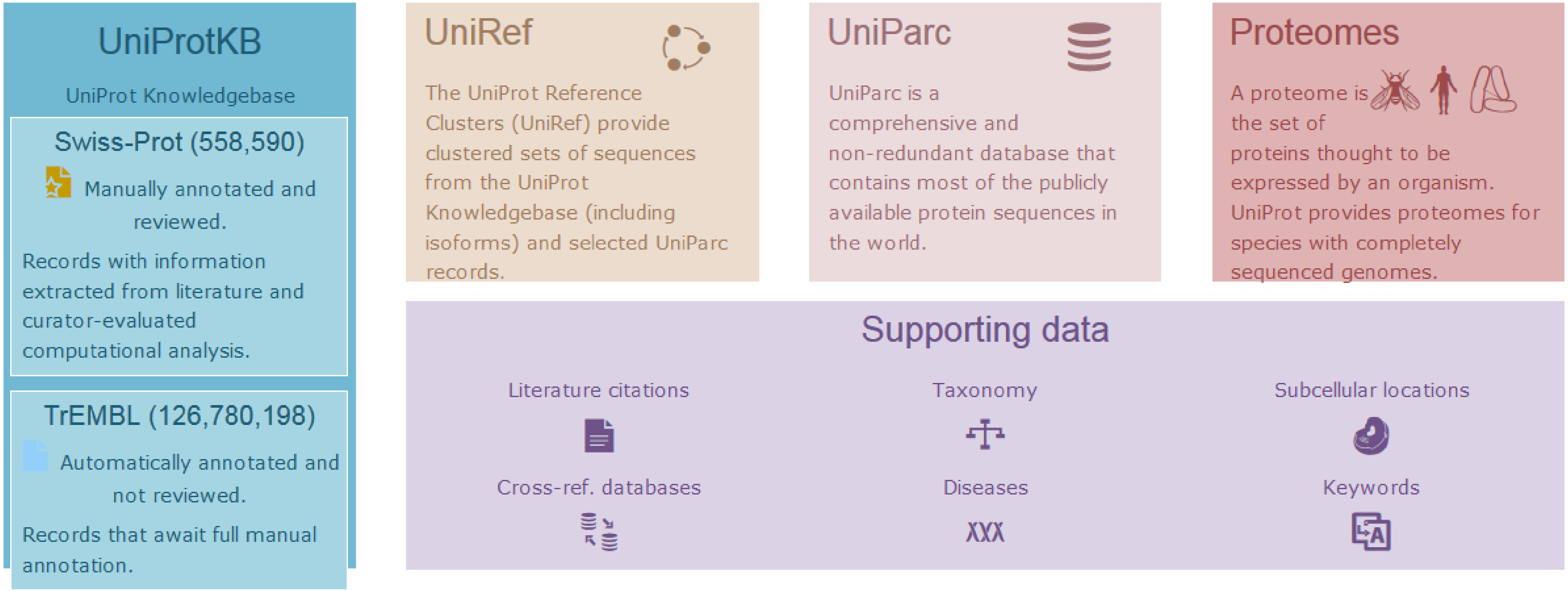
September 2018

There are times in life when things are best left unsaid. So you bite your tongue or someone bites it for you. Either way, you are silenced and no - or less - harm is done. Nature also has its techniques for muffling genes whose products are not necessary at a given time, or that are perhaps harmful once expressed. One technique, which seems to have been with us for a very long time, is DNA methylation...

www.uniprot.org

Swiss Institute of Bioinformatics

UniProt databases



UniProtKB

UniProt Knowledgebase

Swiss-Prot (558,590)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (126,780,198)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

UniProtKB

Each record/entry contains:

Protein sequence(s)
&
Biological knowledge
i.e. function, catalytic activity, subcellular location,..
&
sequence annotation (features)
i.e. PTM, variants, domains, …

Manually annotated or automatically annotated

UniProtKB is composed of 2 sections

UniProtKB/Swiss-Prot

Reviewed - Manually annotated

Records with information extracted from literature and curator-evaluated computational analysis.

One gene / one record

UniProtKB/TrEMBL

Unreviewed – Computationally analyzed

Records that await full manual annotation.

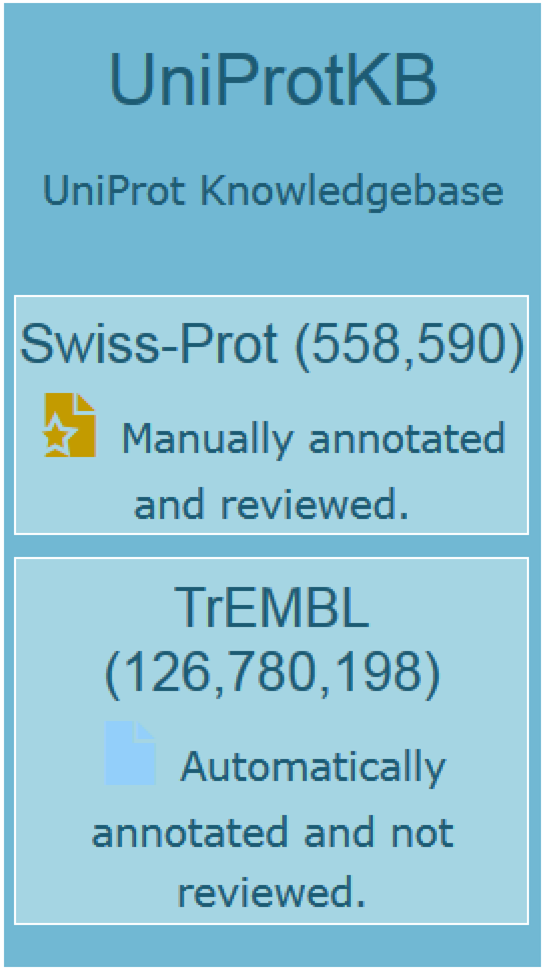
One protein / one record

some redundancy…

released every 4 weeks

UniProtKB is composed of 2 sections

Major differences in the protein sequence and annotation accuracy !



UniProtKB

UniProt Knowledgebase

Swiss-Prot (558,590)

Manually annotated and reviewed.

TrEMBL (126,780,198)

Automatically annotated and not reviewed.

UniProt release 2018_09

0.5 % of UniProtKB protein sequences

99.5 % of UniProtKB protein sequences

UniProtKB: source of protein sequences

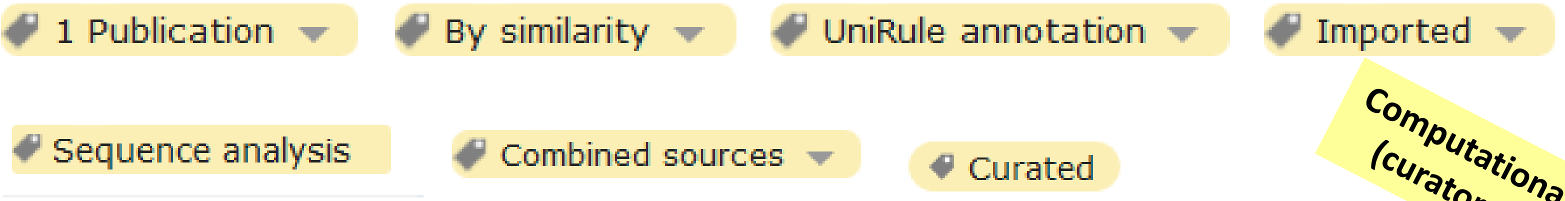
- **INSDC (EMBL/ENA, GenBank, DDBJ) (> 95%)**
- **Ensembl (gene prediction)**
- RefSeq
- Sequences of PDB
- Sequences scanned from literature (PRF)
- Direct submission (including direct protein sequencing)

No evidence nor quality statement for the protein sequence

https://www.uniprot.org/help/sequence_origin

UniProtKB: source of annotation/evidence statements

UniProtKB/Swiss-Prot: Manual insertion, color in yellow



1 Publication

By similarity

UniRule annotation

Imported

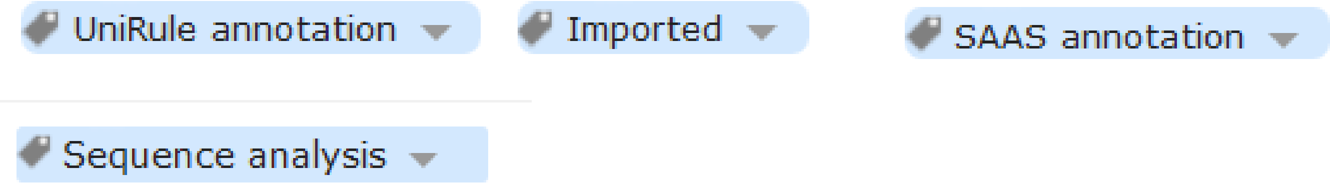
Sequence analysis

Combined sources

Curated

Computational analysis (curator-evaluated)

UniProtKB/TrEMBL: Automated insertion, color in blue



UniRule annotation

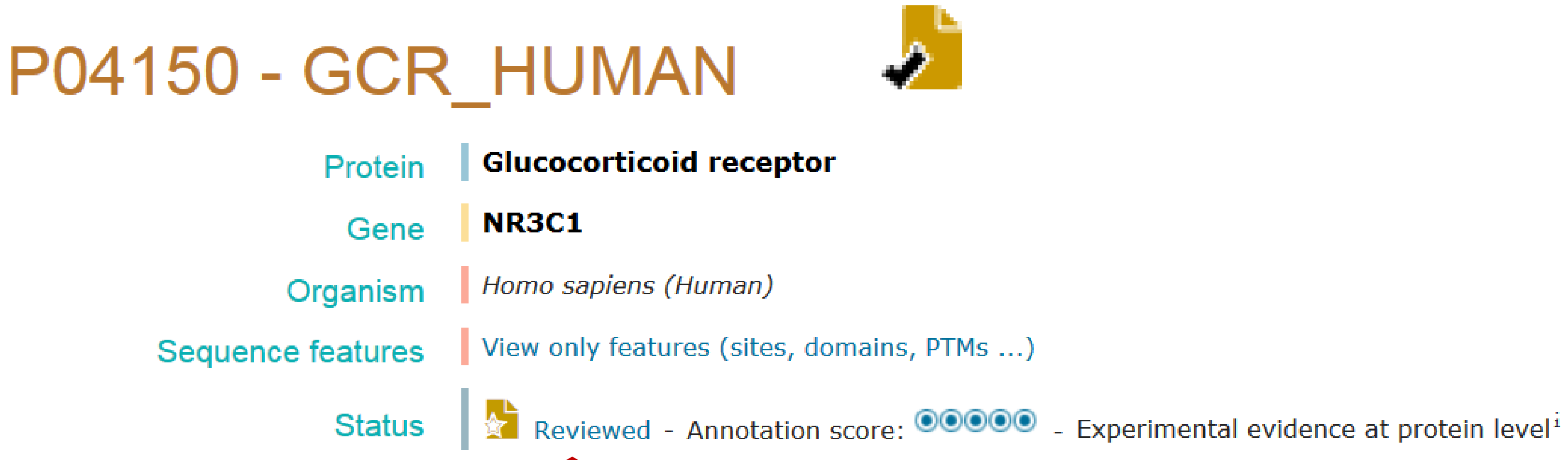
Imported

SAAS annotation

Sequence analysis

Computational analysis (NOT curator-evaluated)

UniProtKB/Swiss-Prot entry



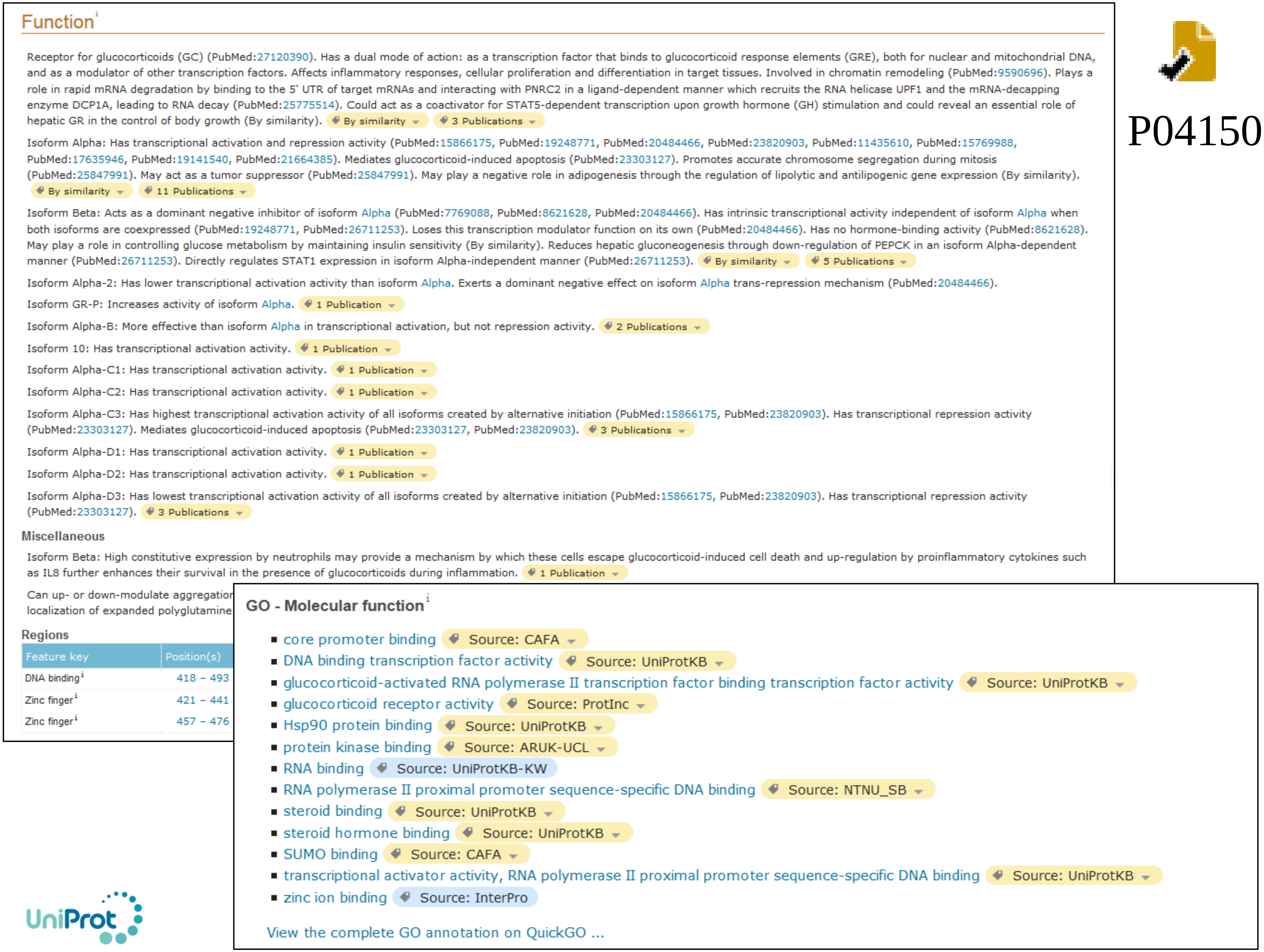
P04150 - GCR_HUMAN

Protein	**Glucocorticoid receptor**
Gene	**NR3C1**
Organism	*Homo sapiens (Human)*
Sequence features	View only features (sites, domains, PTMs ...)
Status	Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

1. Experimental evidence at protein level
2. Experimental evidence at transcript level
3. Protein inferred from homology
4. Protein predicted
5. Protein uncertain

https://www.uniprot.org/help/protein_existence

P04150



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

Isoform Alpha: Has transcriptional activation and repression activity (PubMed:15866175, PubMed:19248771, PubMed:20484466, PubMed:23820903, PubMed:11435610, PubMed:15769988, PubMed:17635946, PubMed:19141540, PubMed:21664385). Mediates glucocorticoid-induced apoptosis (PubMed:23303127). Promotes accurate chromosome segregation during mitosis (PubMed:25847991). May act as a tumor suppressor (PubMed:25847991). May play a negative role in adipogenesis through the regulation of lipolytic and antilipogenic gene expression (By similarity). By similarity 11 Publications

Isoform Beta: Acts as a dominant negative inhibitor of isoform Alpha (PubMed:7769088, PubMed:8621628, PubMed:20484466). Has intrinsic transcriptional activity independent of isoform Alpha when both isoforms are coexpressed (PubMed:19248771, PubMed:26711253). Loses this transcription modulator function on its own (PubMed:20484466). Has no hormone-binding activity (PubMed:8621628). May play a role in controlling glucose metabolism by maintaining insulin sensitivity (By similarity). Reduces hepatic gluconeogenesis through down-regulation of PEPCK in an isoform Alpha-dependent manner (PubMed:26711253). Directly regulates STAT1 expression in isoform Alpha-independent manner (PubMed:26711253). By similarity 5 Publications

Isoform Alpha-2: Has lower transcriptional activation activity than isoform Alpha. Exerts a dominant negative effect on isoform Alpha trans-repression mechanism (PubMed:20484466).

Isoform GR-P: Increases activity of isoform Alpha. 1 Publication

Isoform Alpha-B: More effective than isoform Alpha in transcriptional activation, but not repression activity. 2 Publications

Isoform 10: Has transcriptional activation activity. 1 Publication

Isoform Alpha-C1: Has transcriptional activation activity. 1 Publication

Isoform Alpha-C2: Has transcriptional activation activity. 1 Publication

Isoform Alpha-C3: Has highest transcriptional activation activity of all isoforms created by alternative initiation (PubMed:15866175, PubMed:23820903). Has transcriptional repression activity (PubMed:23303127). Mediates glucocorticoid-induced apoptosis (PubMed:23303127, PubMed:23820903). 3 Publications

Isoform Alpha-D1: Has transcriptional activation activity. 1 Publication

Isoform Alpha-D2: Has transcriptional activation activity. 1 Publication

Isoform Alpha-D3: Has lowest transcriptional activation activity of all isoforms created by alternative initiation (PubMed:15866175, PubMed:23820903). Has transcriptional repression activity (PubMed:23303127). 3 Publications

Miscellaneous

Isoform Beta: High constitutive expression by neutrophils may provide a mechanism by which these cells escape glucocorticoid-induced cell death and up-regulation by proinflammatory cytokines such as IL8 further enhances their survival in the presence of glucocorticoids during inflammation. 1 Publication

Can up- or down-modulate aggregation
localization of expanded polyglutamine

Regions

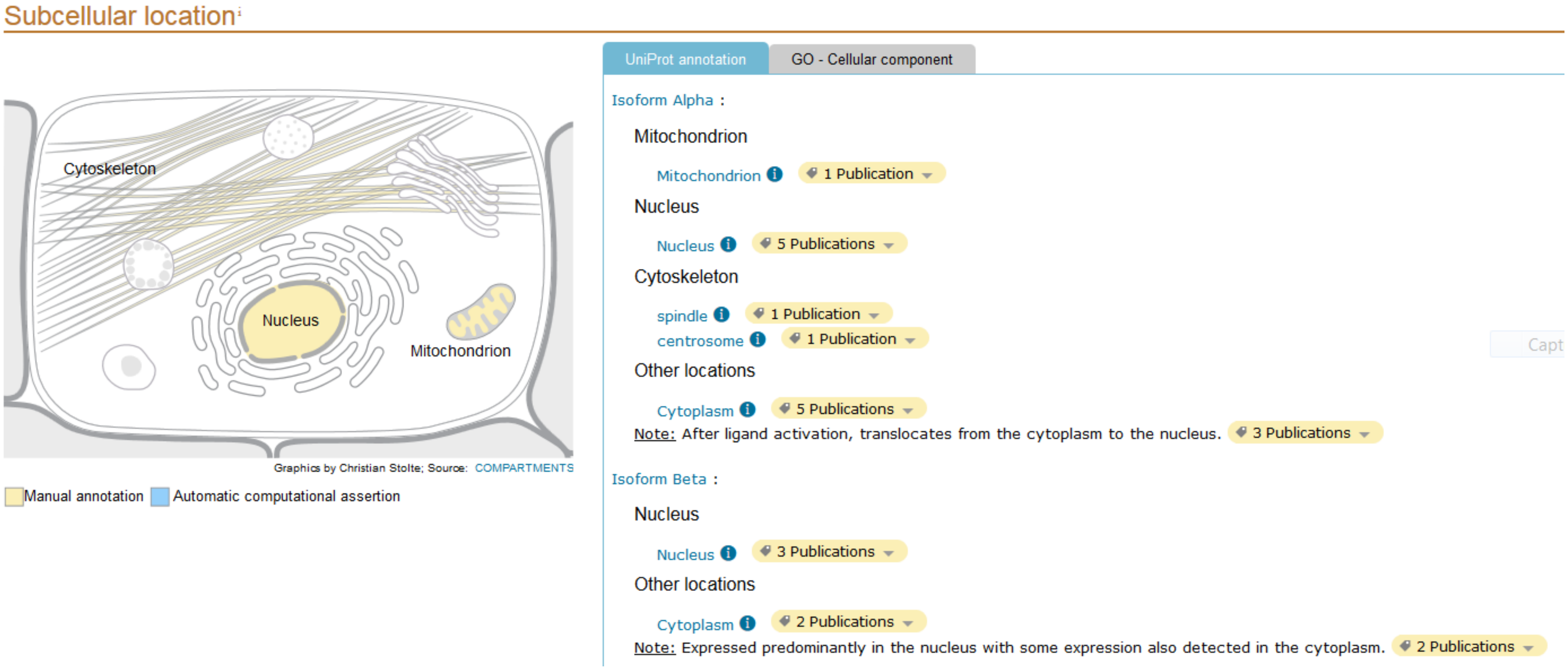
Feature key	Position(s)
DNA binding	418 – 493
Zinc finger	421 – 441
Zinc finger	457 – 476

GO - Molecular function

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

View the complete GO annotation on QuickGO ...

P04150



Subcellular location

UniProt annotation | GO - Cellular component

Graphics by Christian Stolte; Source: COMPARTMENTS

🟨 Manual annotation 🟦 Automatic computational assertion

Isoform Alpha :

Mitochondrion

Mitochondrion 1 Publication

Nucleus

Nucleus 5 Publications

Cytoskeleton

spindle 1 Publication

centrosome 1 Publication

Other locations

Cytoplasm 5 Publications

Note: After ligand activation, translocates from the cytoplasm to the nucleus. 3 Publications

Capt

Isoform Beta :

Nucleus

Nucleus 3 Publications

Other locations

Cytoplasm 2 Publications

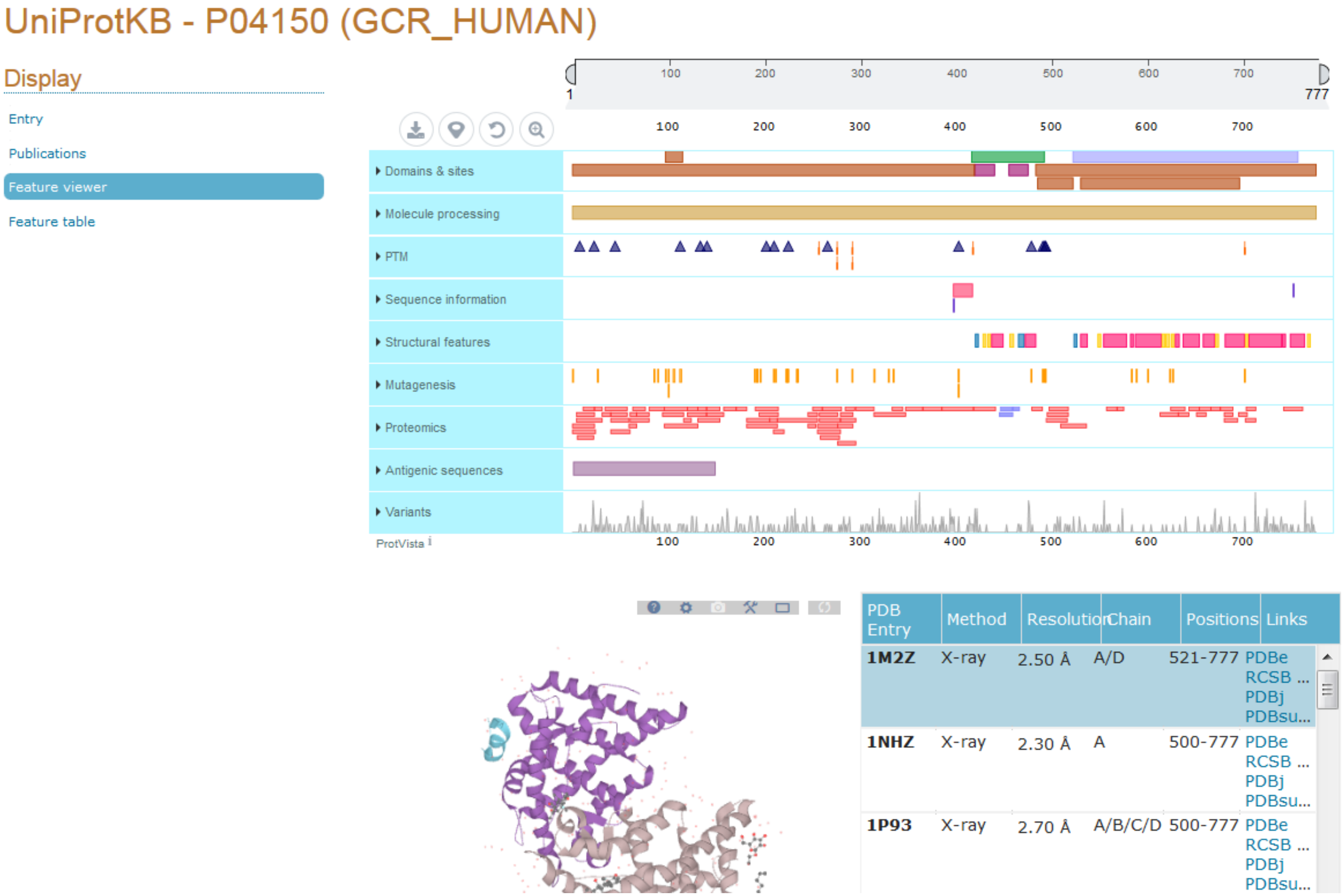
Note: Expressed predominantly in the nucleus with some expression also detected in the cytoplasm. 2 Publications

UniProt

SIB Swiss Institute of Bioinformatics

Sequence annotation (feature viewer)

P04150

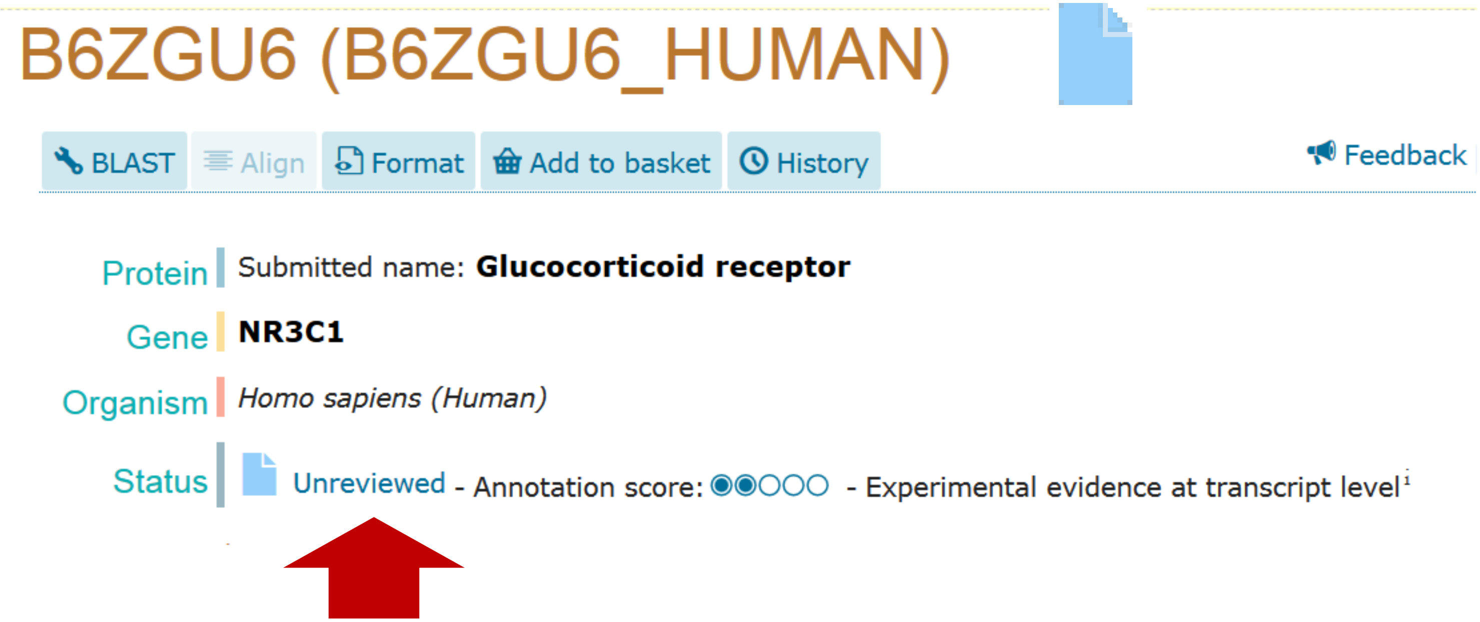


UniProtKB - P04150 (GCR_HUMAN)

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

Swiss Institute of Bioinformatics

UniProtKB/TrEMBL entry



B6ZGU6 (B6ZGU6_HUMAN)

Protein | Submitted name: **Glucocorticoid receptor**

Gene | **NR3C1**

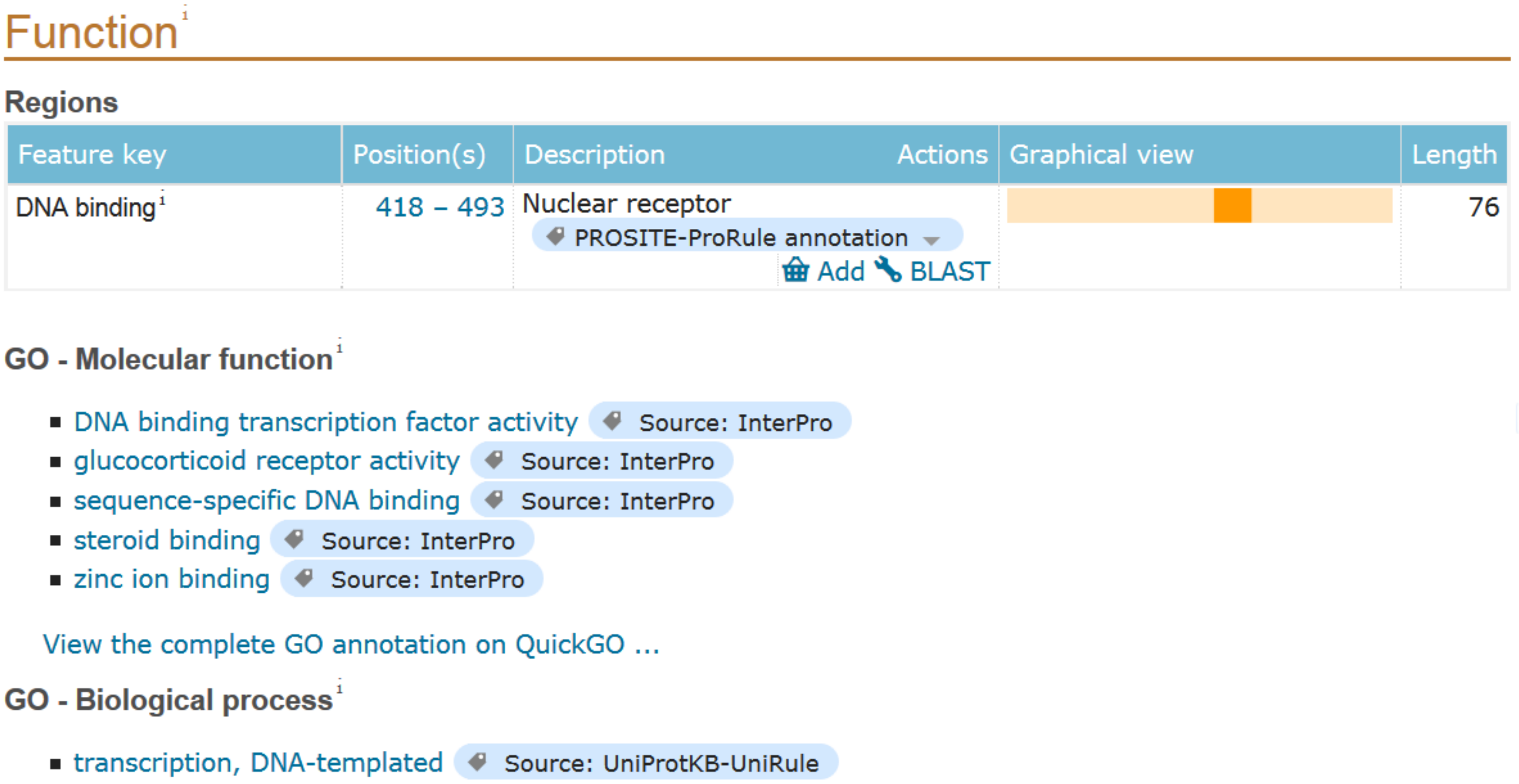
Organism | *Homo sapiens (Human)*

Status | Unreviewed - Annotation score: ●●○○○ - Experimental evidence at transcript level[i]

The same gene (NR3C1) in UniProtKB/TrEMBL

Due to some redundancy between Swiss-Prot and TrEMBL

B6ZGU6



Function

Regions

Feature key	Position(s)	Description	Actions	Graphical view	Length
DNA binding	418 – 493	Nuclear receptor PROSITE-ProRule annotation	Add BLAST		76

GO - Molecular function

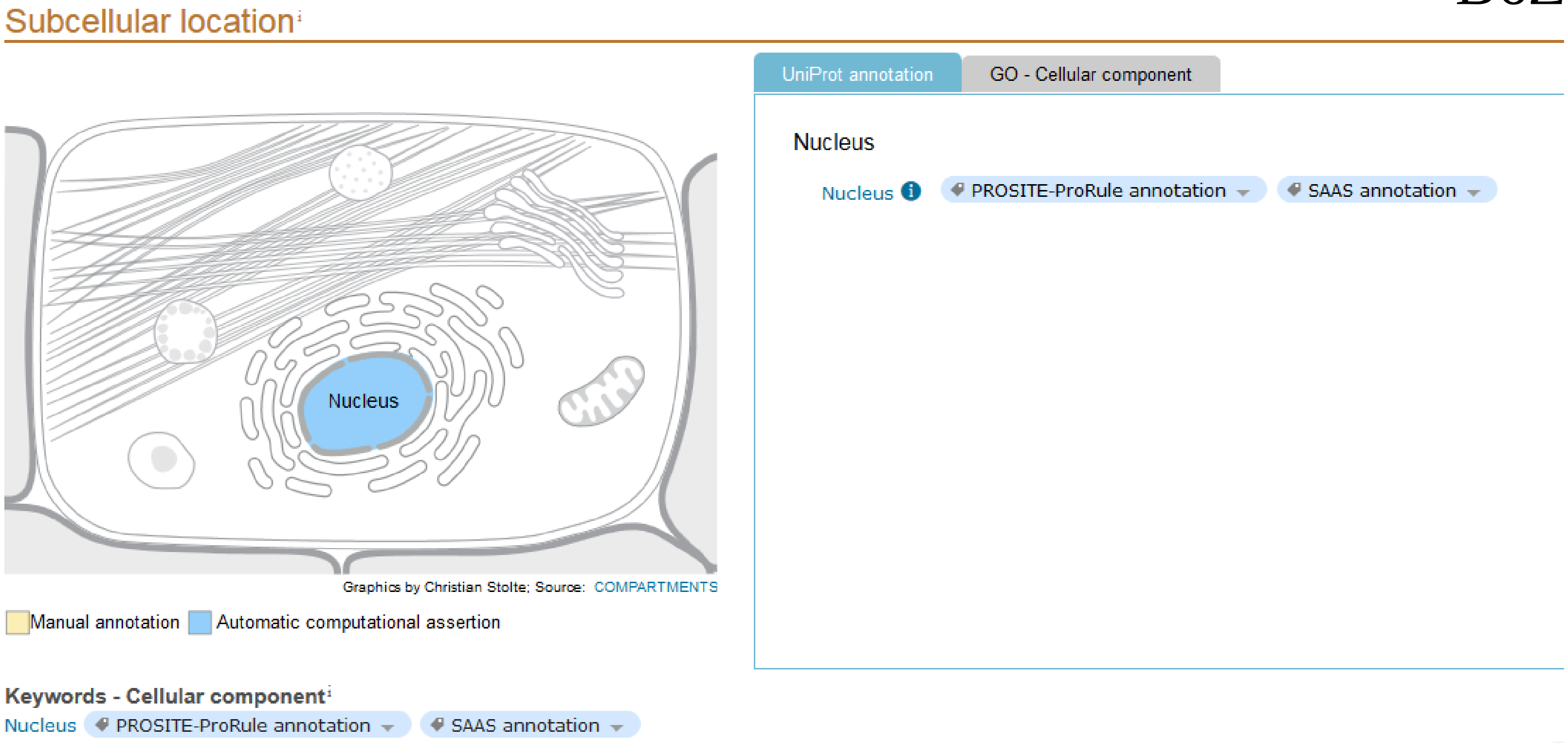
- DNA binding transcription factor activity Source: InterPro
- glucocorticoid receptor activity Source: InterPro
- sequence-specific DNA binding Source: InterPro
- steroid binding Source: InterPro
- zinc ion binding Source: InterPro

View the complete GO annotation on QuickGO ...

GO - Biological process

- transcription, DNA-templated Source: UniProtKB-UniRule

B6ZGU6



Subcellular location

Graphics by Christian Stolte; Source: COMPARTMENTS

🟨 Manual annotation 🟦 Automatic computational assertion

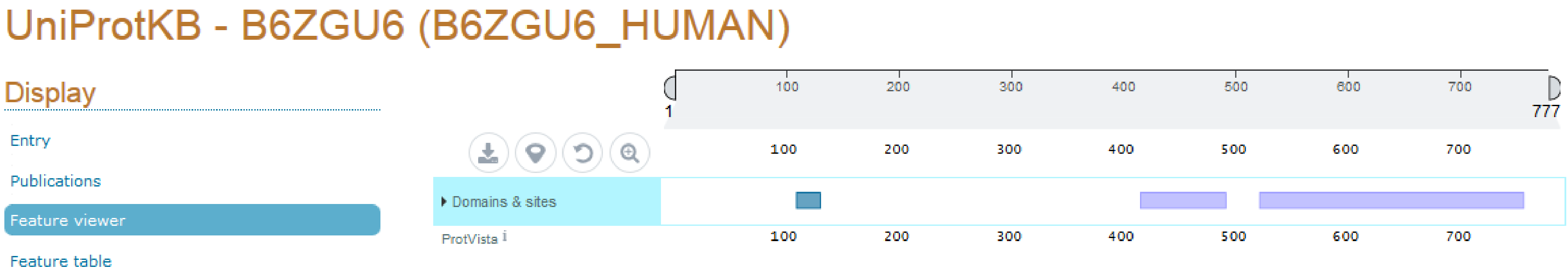
Nucleus

Nucleus ℹ PROSITE-ProRule annotation ▾ SAAS annotation ▾

Keywords - Cellular component

Nucleus PROSITE-ProRule annotation ▾ SAAS annotation ▾

B6ZGU6



UniProtKB - B6ZGU6 (B6ZGU6_HUMAN)

Display

- Entry
- Publications
- Feature viewer
- Feature table

Swiss Institute of Bioinformatics

Nucleic acid sequence databases
- **INSDC, RefSeq, Ensembl**

Protein sequence databases

UniProtKB

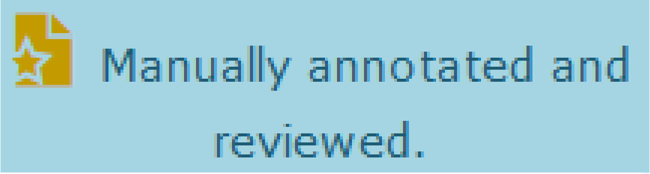
- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals

UniProtKB/Swiss-Prot

Protein sequence



Manually annotated and reviewed.

UniProtKB/Swiss-Prot

⚠ **gene-centric** /protein-centric

- all protein products encoded by one gene are described in a single entry
- One or several protein sequences (isoforms) per entry (canonical / isoform)
- curated sequences in UniProtKB/Swiss-Prot are based on the reference genome (when available)
- collaboration with other databases

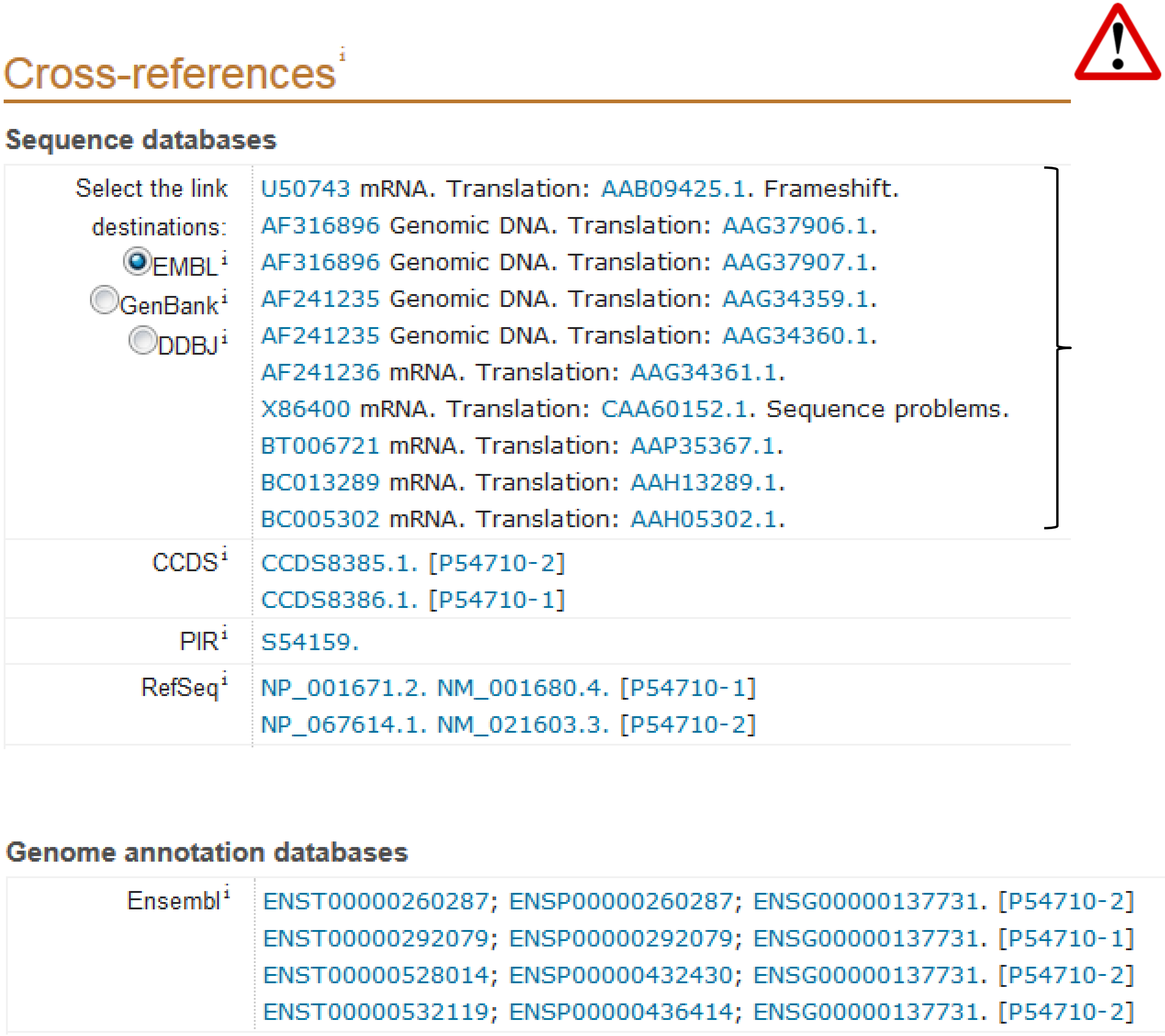
The protein sequences

found in UniProtKB/Swiss-Prot

are often derived from different different nucleotide sequences from EMBL/GenBank/DDBJ entries

+

mapped to the corresponding genomic sequence gene-centric/protein-centric



Cross-references

Sequence databases

Select the link destinations: ● EMBL ○ GenBank ○ DDBJ	U50743 mRNA. Translation: AAB09425.1. Frameshift. AF316896 Genomic DNA. Translation: AAG37906.1. AF316896 Genomic DNA. Translation: AAG37907.1. AF241235 Genomic DNA. Translation: AAG34359.1. AF241235 Genomic DNA. Translation: AAG34360.1. AF241236 mRNA. Translation: AAG34361.1. X86400 mRNA. Translation: CAA60152.1. Sequence problems. BT006721 mRNA. Translation: AAP35367.1. BC013289 mRNA. Translation: AAH13289.1. BC005302 mRNA. Translation: AAH05302.1.
CCDS	CCDS8385.1. [P54710-2] CCDS8386.1. [P54710-1]
PIR	S54159.
RefSeq	NP_001671.2. NM_001680.4. [P54710-1] NP_067614.1. NM_021603.3. [P54710-2]

Genome annotation databases

Ensembl	ENST00000260287; ENSP00000260287; ENSG00000137731. [P54710-2] ENST00000292079; ENSP00000292079; ENSG00000137731. [P54710-1] ENST00000528014; ENSP00000432430; ENSG00000137731. [P54710-2] ENST00000532119; ENSP00000436414; ENSG00000137731. [P54710-2]

Used to construct the UniProtKB canonical sequence

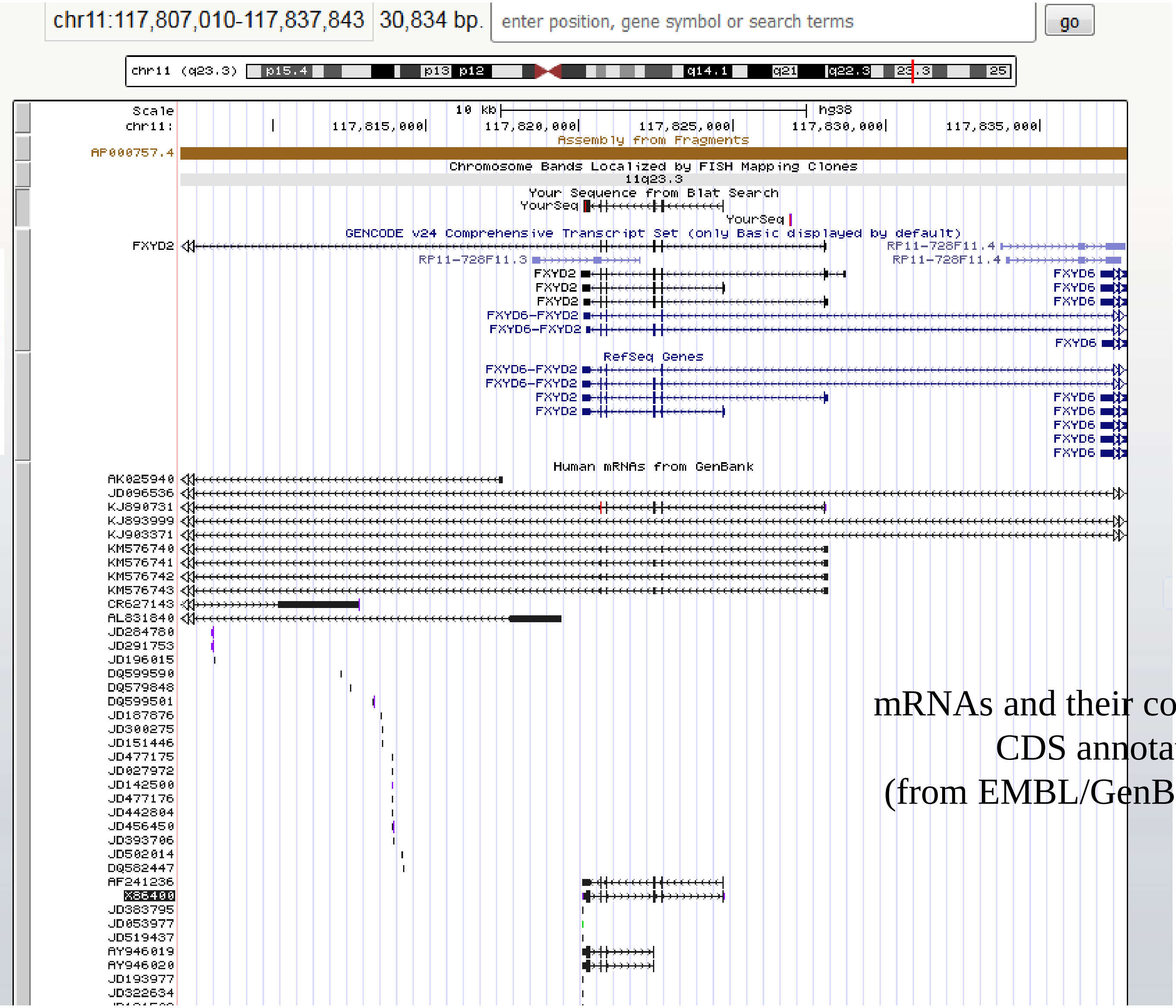
Automatically mapped to the UniProtKB record

http://www.uniprot.org/uniprot/P54710#cross_references

- At least 20% of UniProtKB/Swiss-Prot entries required curation effort to “correct” the sequences.

- Typical problems
  - unsolved conflicts;
  - uncorrected initiation sites;
  - frameshifts;
  - other ‘problems’

UCSC genome browser: another gene

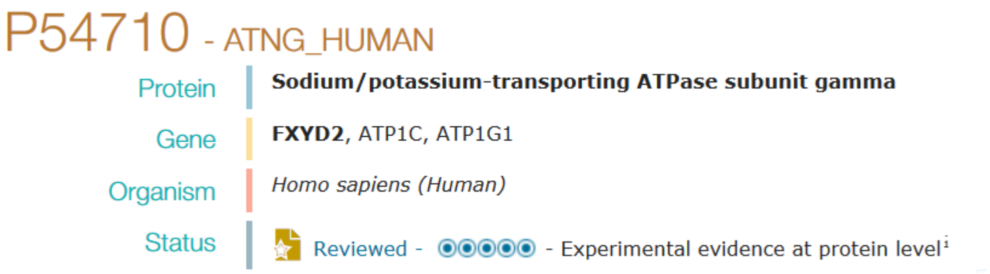


chr11:117,807,010-117,837,843 30,834 bp. enter position, gene symbol or search terms go

mRNAs and their corresponding CDS annotation (from EMBL/GenBank/DDBJ)

UniProt

Swiss Institute of Bioinformatics



P54710 - ATNG_HUMAN

Protein | **Sodium/potassium-transporting ATPase subunit gamma**

Gene | **FXYD2**, ATP1C, ATP1G1

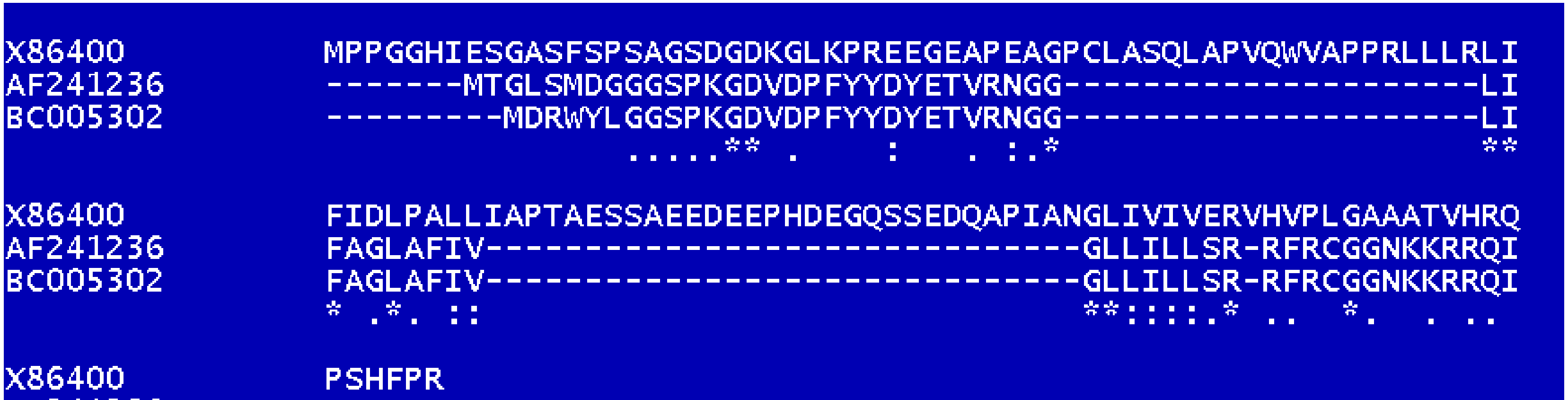
Organism | *Homo sapiens (Human)*

Status | Reviewed - ●●●●● - Experimental evidence at protein level[i]

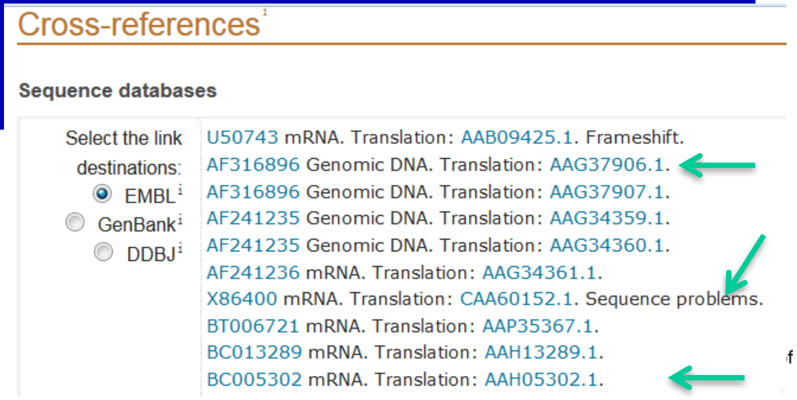
Manually checked (comparison with orthologs, etc.)

incorrect

correct



```
X86400          MPPGGHIESGASFSPSAGSDGDKGLKPREEGEAPEAGPCLASQLAPVQWVAPPRLLLRLI
AF241236        -------MTGLSMDGGGSPKGDVDPFYYDYETVRNGG---------------------LI
BC005302        ---------MDRWYLGGSPKGDVDPFYYDYETVRNGG---------------------LI
                               .....** .    :   . :.*                     **

X86400          FIDLPALLIAPTAESSAEEDEEPHDEGQSSEDQAPIANGLIVIVERVHVPLGAAATVHRQ
AF241236        FAGLAFIV------------------------------GLLILLSR-RFRCGGNKKRRQI
BC005302        FAGLAFIV------------------------------GLLILLSR-RFRCGGNKKRRQI
                * .*. ::                              **::::.* ..  *.  . ..

X86400          PSHFPR
AF241236        NEDEP-
BC005302        NEDEP-
                 .  *
```



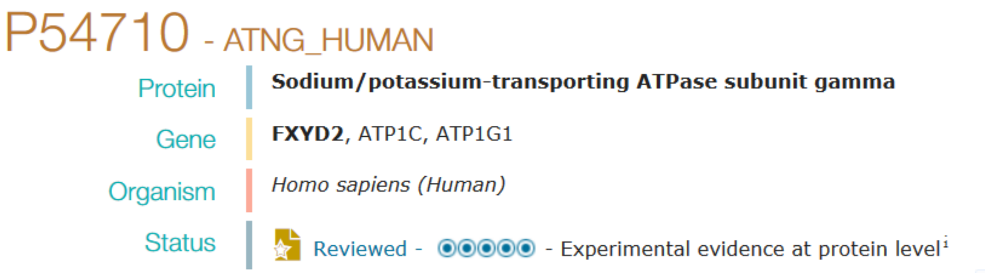
Cross-references[i]

Sequence databases

Select the link destinations:
● EMBL[i]
○ GenBank[i]
○ DDBJ[i]

U50743 mRNA. Translation: AAB09425.1. Frameshift.
AF316896 Genomic DNA. Translation: AAG37906.1.
AF316896 Genomic DNA. Translation: AAG37907.1.
AF241235 Genomic DNA. Translation: AAG34359.1.
AF241235 Genomic DNA. Translation: AAG34360.1.
AF241236 mRNA. Translation: AAG34361.1.
X86400 mRNA. Translation: CAA60152.1. Sequence problems.
BT006721 mRNA. Translation: AAP35367.1.
BC013289 mRNA. Translation: AAH13289.1.
BC005302 mRNA. Translation: AAH05302.1.

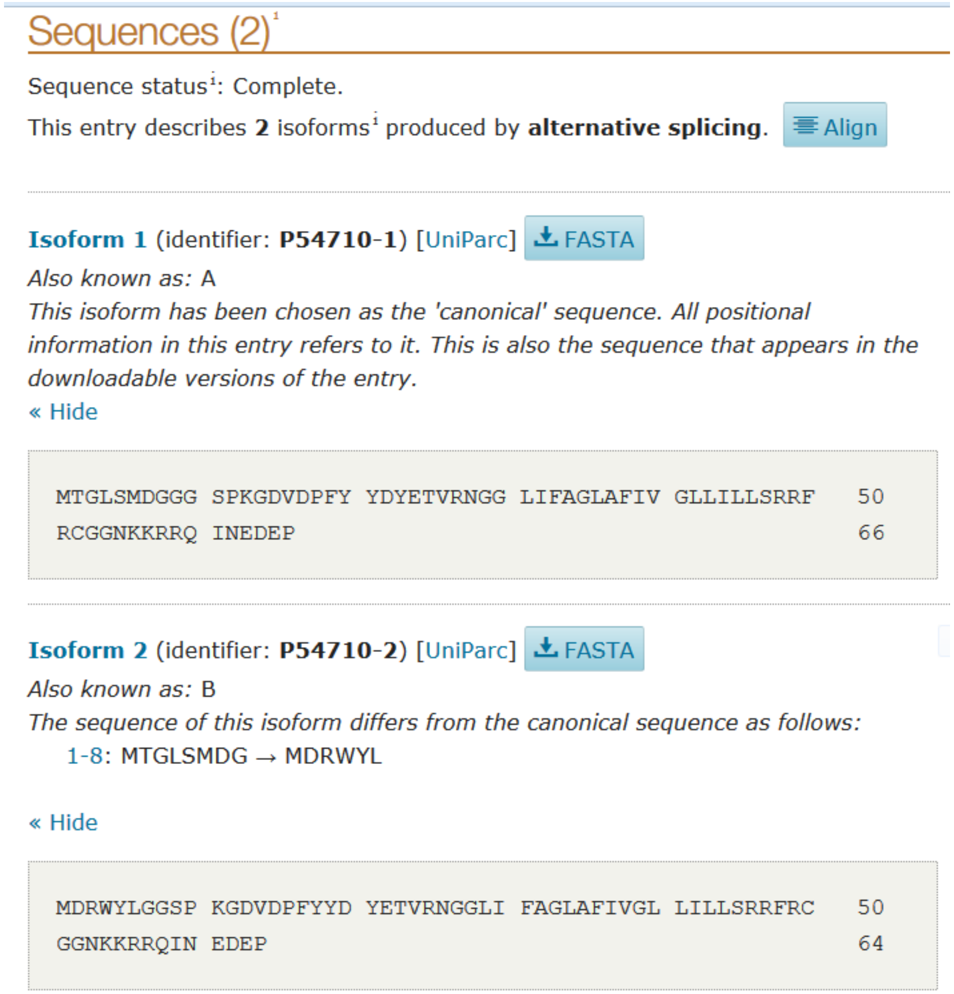
UniProt



P54710 - ATNG_HUMAN

Protein	**Sodium/potassium-transporting ATPase subunit gamma**
Gene	**FXYD2**, ATP1C, ATP1G1
Organism	*Homo sapiens (Human)*
Status	Reviewed - ●●●●● - Experimental evidence at protein level[i]

gene-centric / protein-centric



Sequences (2)[i]

Sequence status[i]: Complete.

This entry describes **2** isoforms[i] produced by **alternative splicing**. Align

Isoform 1 (identifier: **P54710-1**) [UniParc] FASTA

Also known as: A

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

```
MTGLSMDGGG SPKGDVDPFY YDYETVRNGG LIFAGLAFIV GLLILLSRRF    50
RCGGNKKRRQ INEDEP                                         66
```

Isoform 2 (identifier: **P54710-2**) [UniParc] FASTA

Also known as: B

The sequence of this isoform differs from the canonical sequence as follows:

1-8: MTGLSMDG → MDRWYL

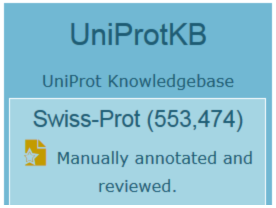
« Hide

```
MDRWYLGGSP KGDVDPFYYD YETVRNGGLI FAGLAFIVGL LILLSRRFRC    50
GGNKKRRQIN EDEP                                           64
```

UniProt

SIB Swiss Institute of Bioinformatics

UniProtKB/Swiss-Prot


UniProtKB

UniProt Knowledgebase

Swiss-Prot (553,474)

Manually annotated and reviewed.

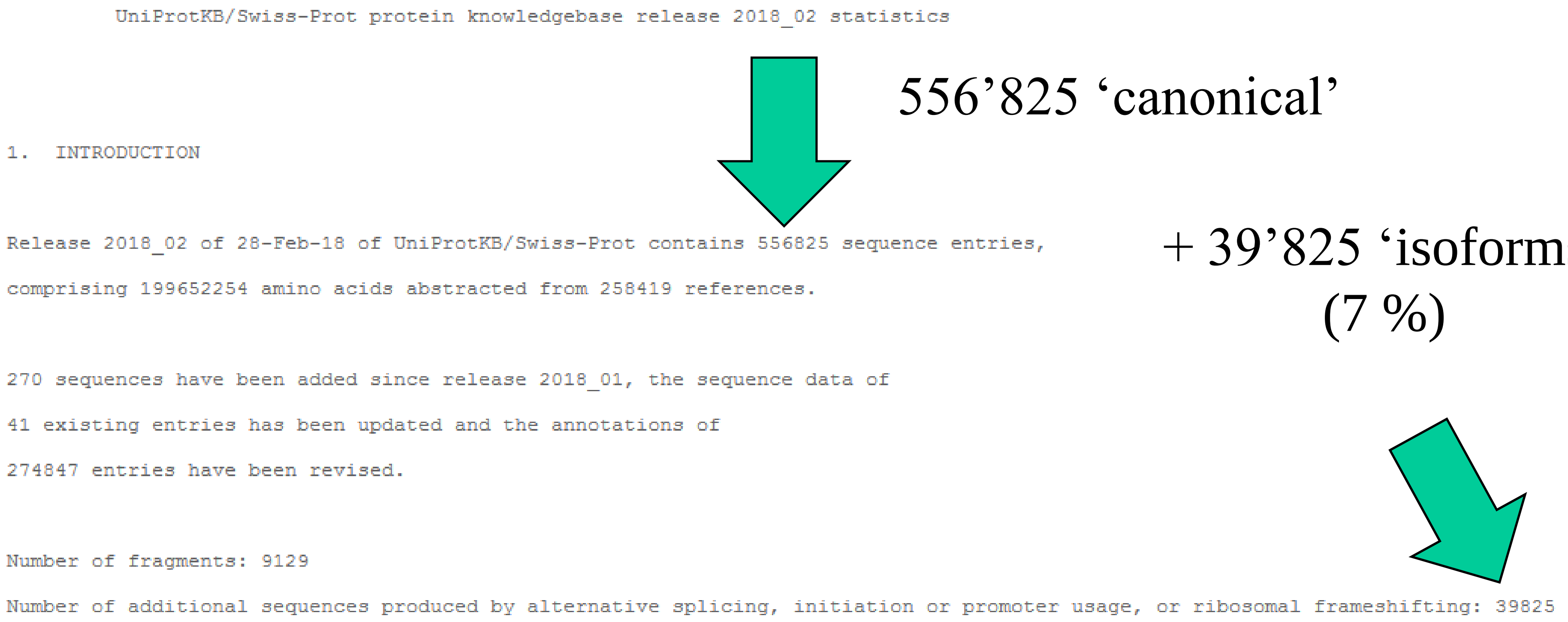
Entry vs Protein sequence(s)

One entry – one gene – one species

One or several protein sequences (isoforms) per entry

canonical & isoform

⚠ **gene-centric** / protein-centric



```
        UniProtKB/Swiss-Prot protein knowledgebase release 2018_02 statistics


1.   INTRODUCTION


Release 2018_02 of 28-Feb-18 of UniProtKB/Swiss-Prot contains 556825 sequence entries,

comprising 199652254 amino acids abstracted from 258419 references.


270 sequences have been added since release 2018_01, the sequence data of

41 existing entries has been updated and the annotations of

274847 entries have been revised.


Number of fragments: 9129

Number of additional sequences produced by alternative splicing, initiation or promoter usage, or ribosomal frameshifting: 39825
```

556’825 ‘canonical’

+ 39’825 ‘isoforms’
(7 %)

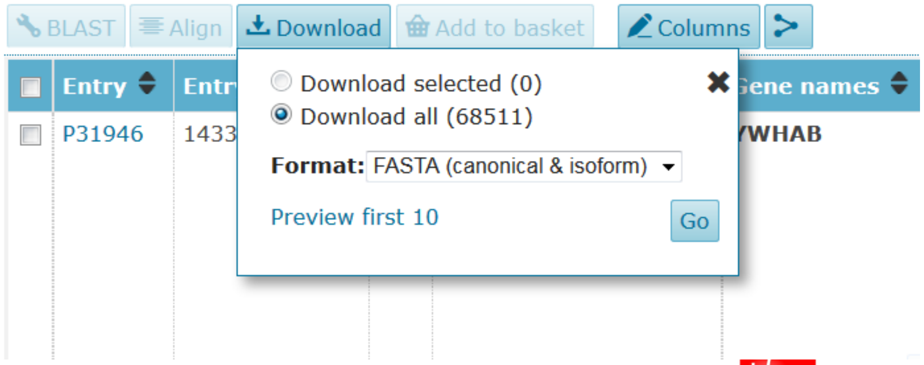
http://web.expasy.org/docs/relnotes/relstat.html

Beware

The isoform sequences are not included in all datasets,

Examples:

- Complete proteome -> download Fasta (canonical & isoform)
- Blast@ NCBI (NCBInr)

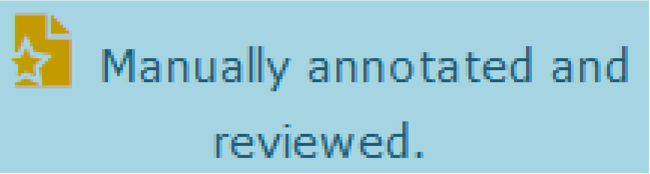


○ Download selected (0)
● Download all (68511)

Format: FASTA (canonical & isoform)

UniProtKB/Swiss-Prot

Biological knowledge / annotation



Manually annotated and reviewed.

Knowledge:

- comprehensive summary (free text) that provides a complete overview of the information available
- standardized vocabularies to facilitate subsequent retrieval whenever possible

Source of annotation/Evidence statements

- Selected Publication (experimental)

1 Publication

Curated

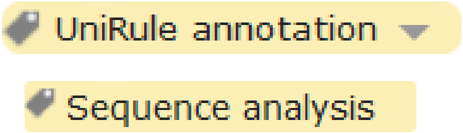
- Another UniProtKB entry (orthologs):

By similarity

- An entry from another database:

Imported

- Curator-evaluated computational analysis*



UniRule annotation

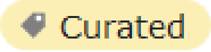
Sequence analysis

- Combined sources

Combined sources

** more details later…*

Source of annotation/Evidence statements

- Selected Publication (experimental)
1 Publication

Curated

- Another UniProtKB entry (orthologs):
By similarity

- An entry from another database:
Imported

- Curator-evaluated computational analysis*
UniRule annotation

Sequence analysis

- Combined sources
Combined sources

** more details later…*

comprehensive and computer friendly representation of biological knowledge

PubMed=16595657

Selected Publication

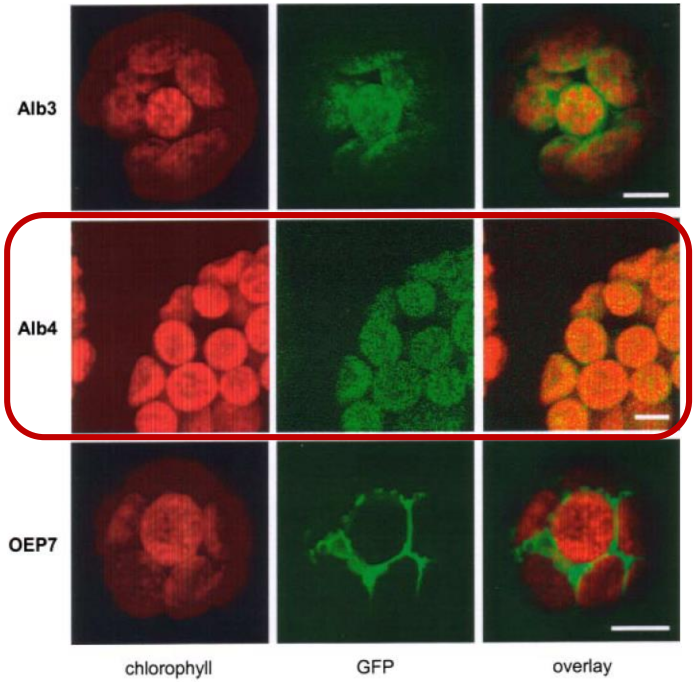
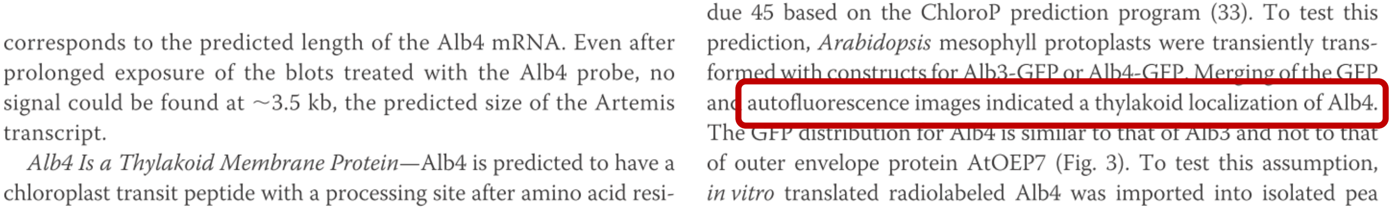


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a



corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript.

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid residue 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

UniProt

SIB
Swiss Institute of Bioinformatics

comprehensive and computer friendly representation of biological knowledge

PubMed=16595657

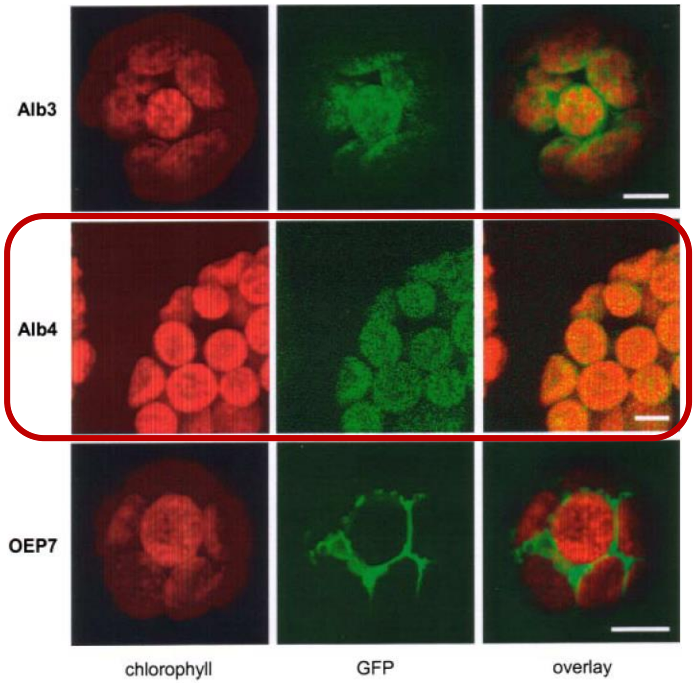
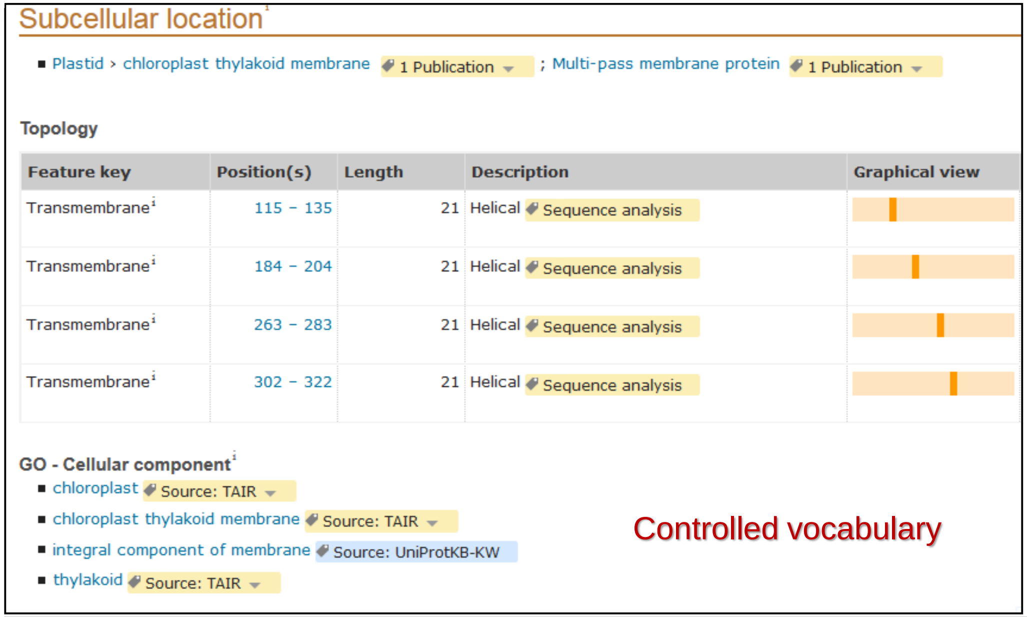


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a

UniProtKB Q9FYL3



Subcellular location

- Plastid › chloroplast thylakoid membrane 1 Publication ; Multi-pass membrane protein 1 Publication

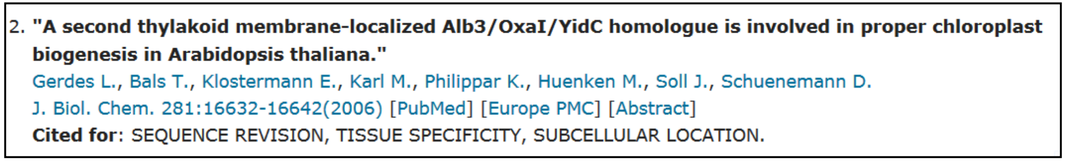
Topology

Feature key	Position(s)	Length	Description	Graphical view
Transmembrane	115 – 135	21	Helical Sequence analysis	
Transmembrane	184 – 204	21	Helical Sequence analysis	
Transmembrane	263 – 283	21	Helical Sequence analysis	
Transmembrane	302 – 322	21	Helical Sequence analysis	

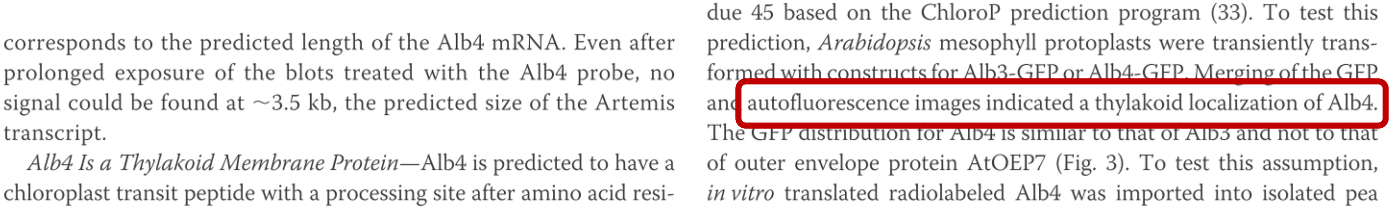
GO - Cellular component

- chloroplast Source: TAIR
- chloroplast thylakoid membrane Source: TAIR
- integral component of membrane Source: UniProtKB-KW
- thylakoid Source: TAIR

Controlled vocabulary



2. **"A second thylakoid membrane-localized Alb3/OxaI/YidC homologue is involved in proper chloroplast biogenesis in Arabidopsis thaliana."**
Gerdes L., Bals T., Klostermann E., Karl M., Philippar K., Huenken M., Soll J., Schuenemann D.
J. Biol. Chem. 281:16632-16642(2006) [PubMed] [Europe PMC] [Abstract]
Cited for: SEQUENCE REVISION, TISSUE SPECIFICITY, SUBCELLULAR LOCATION.



corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript.

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid residue 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

Swiss Institute of Bioinformatics

Source of annotation/Evidence statements

- Selected Publication (experimental)

1 Publication

- Another UniProtKB entry (orthologs):

By similarity

- An entry from another database:

Imported

- Curator-evaluated computational analysis

UniRule annotation

*

Sequence analysis

- Combined sources

Combined sources

** more details later…*

comprehensive and computer friendly representation of biological knowledge

Selected Publication

PubMed=16595657

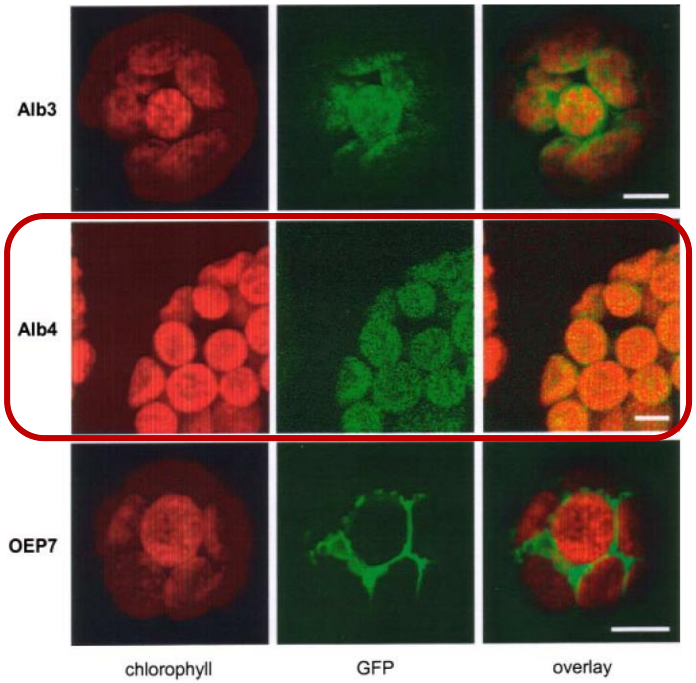
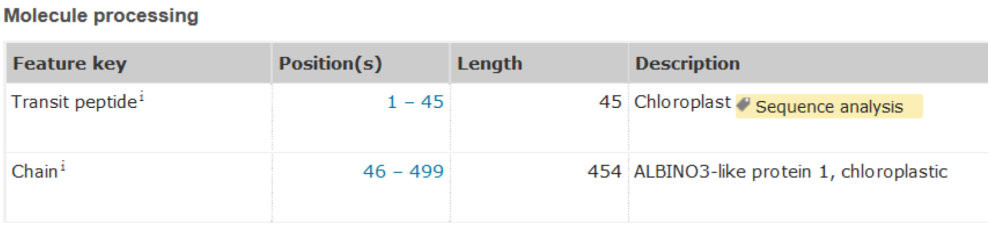


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a

UniProtKB Q9FYL3



Molecule processing

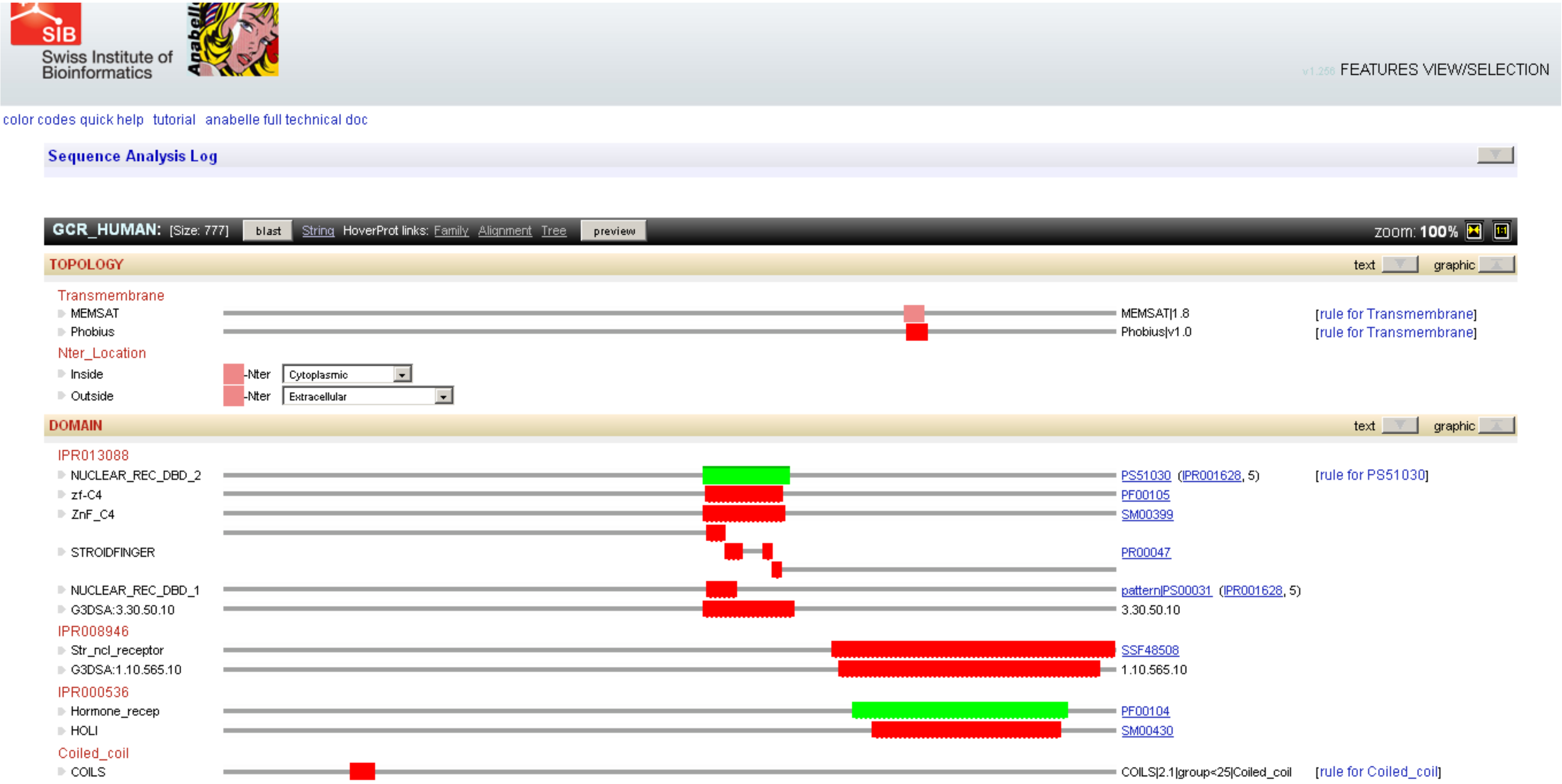
Feature key	Position(s)	Length	Description
Transit peptide[i]	1 – 45	45	Chloroplast Sequence analysis
Chain[i]	46 – 499	454	ALBINO3-like protein 1, chloroplastic

corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript.

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid residue 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

Swiss Institute of Bioinformatics

Protein sequence analysis: in-house resource



v1.256 FEATURES VIEW/SELECTION

color codes quick help tutorial anabelle full technical doc

Sequence Analysis Log

GCR_HUMAN: [Size: 777] blast String HoverProt links: Family Alignment Tree preview zoom: 100%

TOPOLOGY text graphic

Transmembrane
- MEMSAT — MEMSAT|1.8 [rule for Transmembrane]
- Phobius — Phobius|v1.0 [rule for Transmembrane]

Nter_Location
- Inside — -Nter Cytoplasmic
- Outside — -Nter Extracellular

DOMAIN text graphic

IPR013088
- NUCLEAR_REC_DBD_2 — PS51030 (IPR001628, 5) [rule for PS51030]
- zf-C4 — PF00105
- ZnF_C4 — SM00399
- STROIDFINGER — PR00047
- NUCLEAR_REC_DBD_1 — pattern|PS00031 (IPR001628, 5)
- G3DSA:3.30.50.10 — 3.30.50.10

IPR008946
- Str_ncl_receptor — SSF48508
- G3DSA:1.10.565.10 — 1.10.565.10

IPR000536
- Hormone_recep — PF00104
- HOLI — SM00430

Coiled_coil
- COILS — COILS|2.1|group<25|Coiled_coil [rule for Coiled_coil]

Table 1. Sequence analysis tools used during the UniProtKB manual curation process

Program	Version	Prediction
Topology		
Signal P (6)	3.0	Presence and location of signal peptides
TargetP (6)	1.1	Presence and location of transit peptides
Predotar (7)	1.03	Mitochondrial, plastid or ER targeting sequences
ESKW* (8)	UniProt-modified version 1.0	Transmembrane domains
MEMSAT (9)	UniProt-modified version 1.8a	Transmembrane domains
TMHMM (10)	2.0	Transmembrane domains
Phobius (11)	Unknown	Discriminates transmembrane and signal regions
Post-translational modifications		
GPI-predictor (12)	1.0	GPI lipid anchor sites
NetNGlyc (13)	1.0	N-glycosylation sites
NetOGlyc (14)	3.1	O-glycosylation sites
NMT Predictor (15)	1.0	N-terminal myristoylation sites
Sulfinator (16)	1.0	Tyrosine sulfation sites
Domains		
ps_scan	1.0	Internal PROSITE profile, pattern and rule scanning program
InterPro (17)	Uses latest versions of InterPro and InterProScan	Retrieves non-PROSITE motif matches using InterPro database or InterProScan
Coils (18)	2.2	Coiled-coil regions
polyAA	1.0	Internal program which identifies homopolymeric stretches of amino acids
REPEAT (19)	1.1	Identifies the following repeats: Ankyrin, Armadillo, HAT, HEAT, Kelch, Leucine-rich, PFTA, PFTB, RCC1, TPR, WD40

*ESKW = transmembrane prediction algorithm by Eisenberg, Schwarz, Komaromy and Wall

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

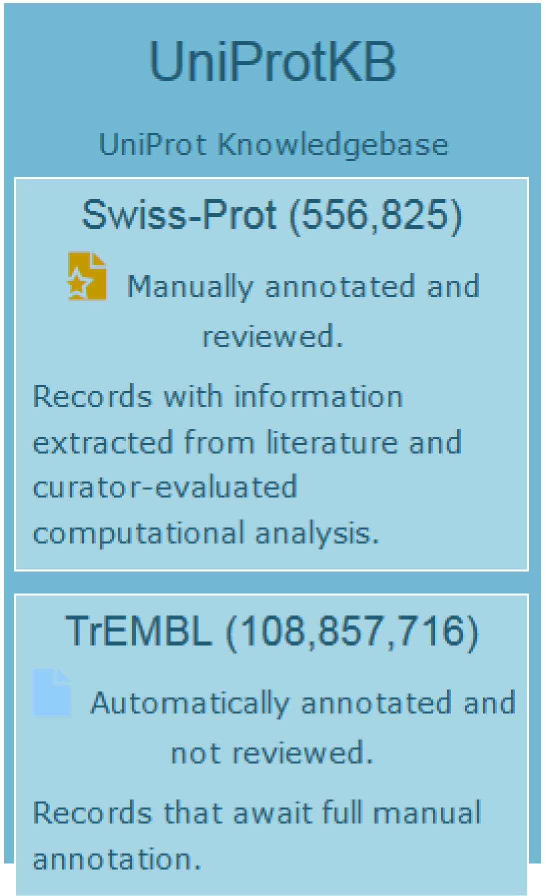
UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals

UniProtKB/TrEMBL



UniProtKB

UniProt Knowledgebase

Swiss-Prot (556,825)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (108,857,716)

Automatically annotated and not reviewed.

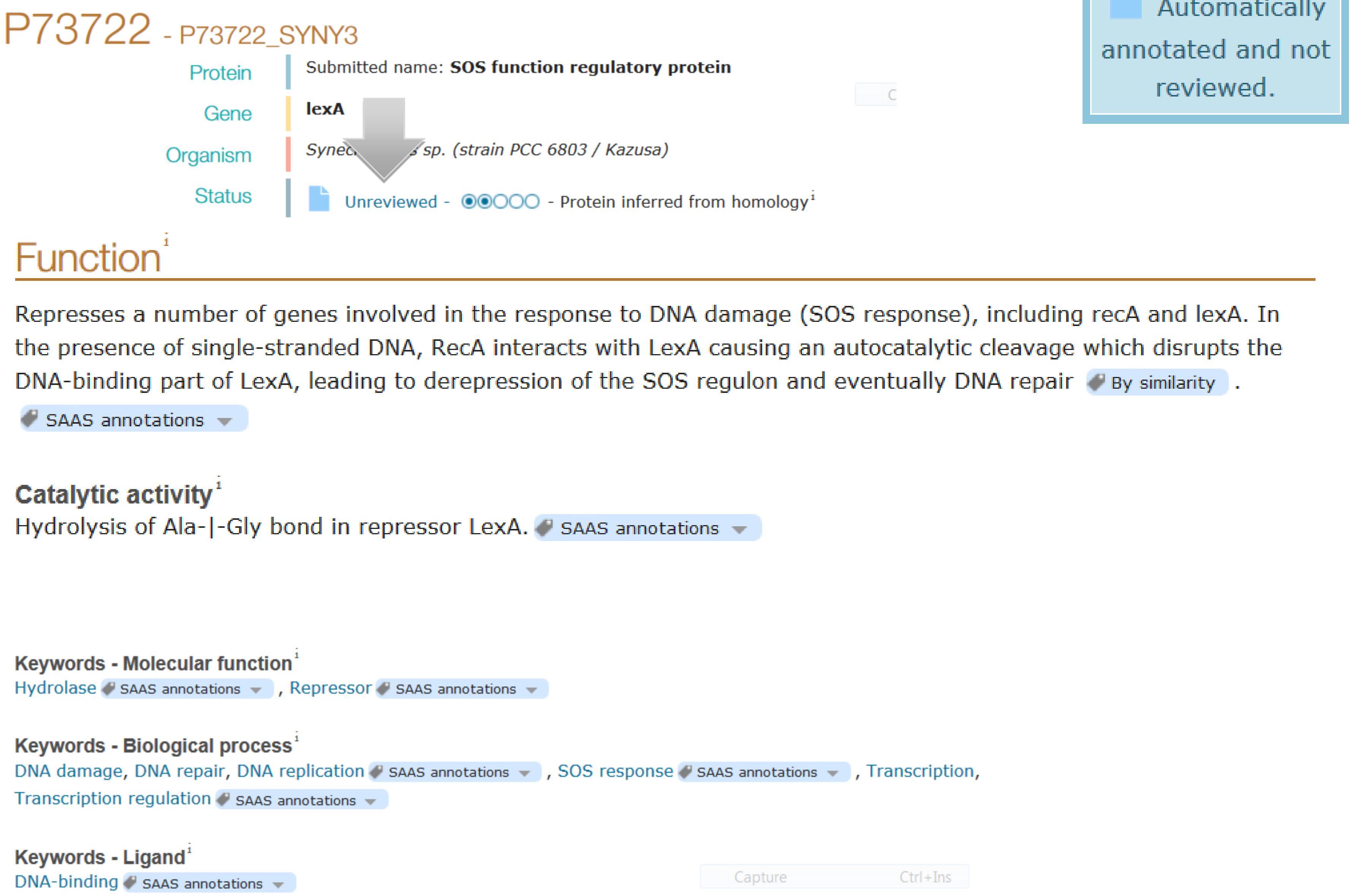
Records that await full manual annotation.

One protein sequence per entry

gene-centric **/protein-centric**

99,5 % of UniProtKB protein sequences

Automatically annotated and not reviewed.



P73722 - P73722_SYNY3

Protein | Submitted name: **SOS function regulatory protein**

Gene | **lexA**

Organism | *Synec s sp. (strain PCC 6803 / Kazusa)*

Status | Unreviewed - ◉◉○○○ - Protein inferred from homology

Function

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 By similarity .

SAAS annotations

Catalytic activity

Hydrolysis of Ala-|-Gly bond in repressor LexA. SAAS annotations

Keywords - Molecular function

Hydrolase SAAS annotations , Repressor SAAS annotations

Keywords - Biological process

DNA damage, DNA repair, DNA replication SAAS annotations , SOS response SAAS annotations , Transcription, Transcription regulation SAAS annotations

Keywords - Ligand

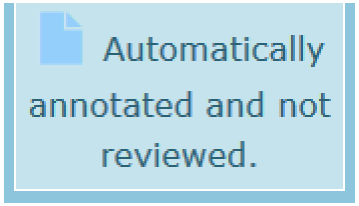
DNA-binding SAAS annotations

UniProt

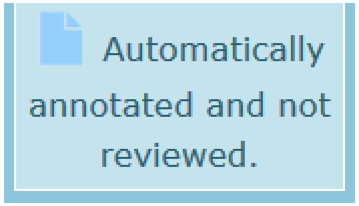
SIB Swiss Institute of Bioinformatics

UniProtKB/TrEMBL

Protein sequence



Automatically annotated and not reviewed.



Automatically annotated and not reviewed.

Protein sequence

- The quality of the protein sequences is dependent on the information provided by the submitter of the original nucleotide entry (EMBL-ENA CDS) or of the gene prediction pipeline (i.e. Ensembl).

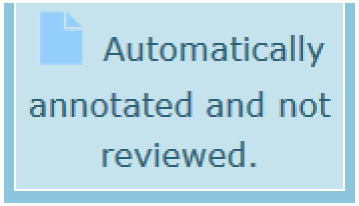
- 100% identical sequences (same length, same organism are merged automatically).

gene-centric **/protein-centric**

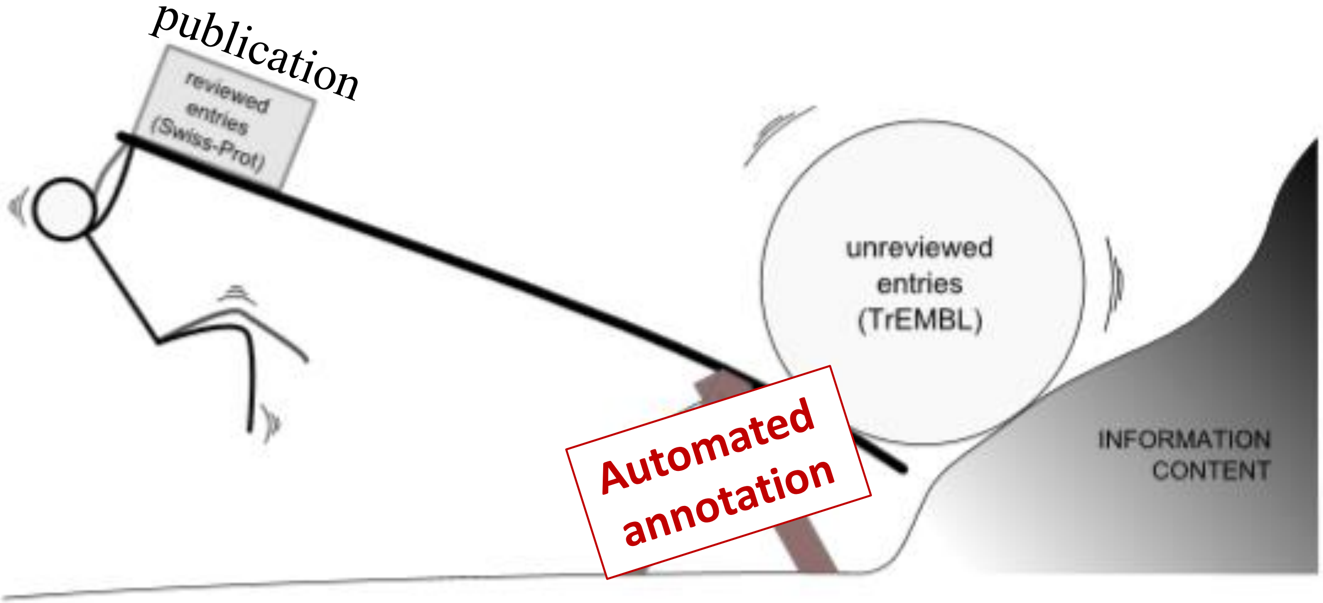
One protein sequence per entry

UniProtKB/TrEMBL

Biological knowledge / annotation

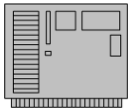


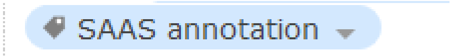
Automatically annotated and not reviewed.


publication
reviewed entries (Swiss-Prot)
unreviewed entries (TrEMBL)
Automated annotation
INFORMATION CONTENT

Automated annotation

Automated generated rules (SAAS)



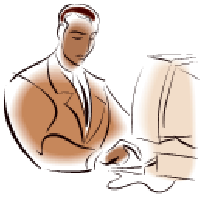


SAAS annotation

Generates a set of decision trees using data mining (new set every UniProtKB release)

Manually generated rules (UniRule)

Maintains a set of manual annotation rules

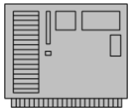


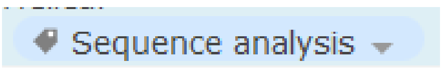
UniRule annotation

UniRule = PIR + HAMAP + Rulebase

Sequence analysis methods (SAM)

Signal, transmembrane, coils prediction

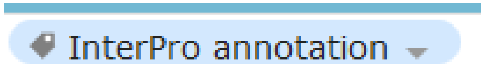




Sequence analysis

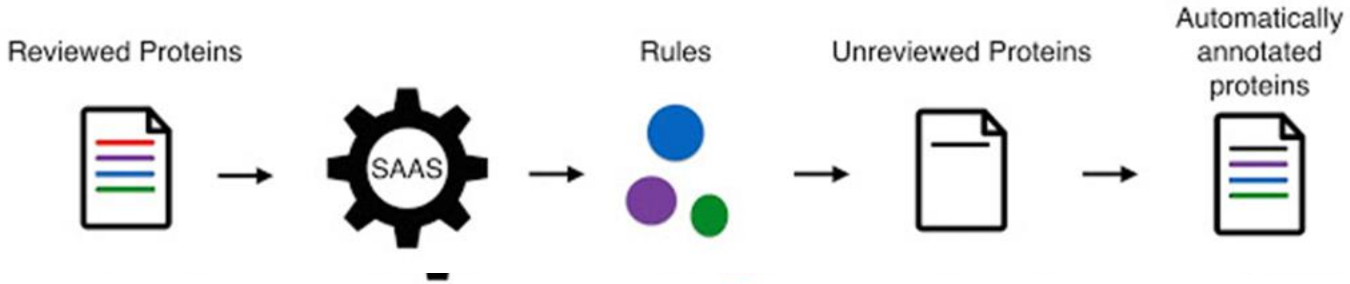
InterPro

Domains & GO terms

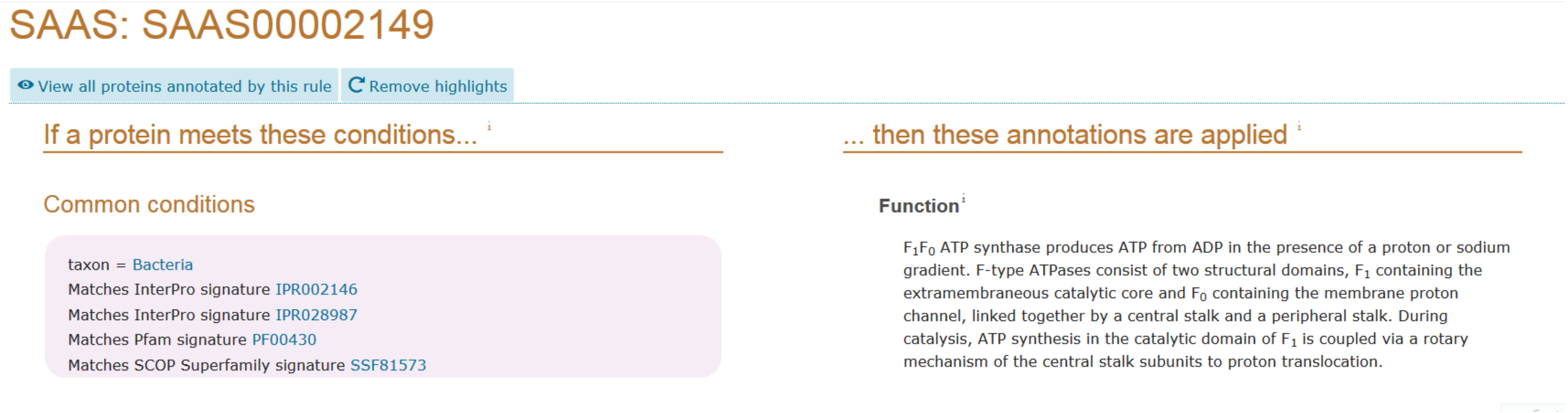


InterPro annotation

SAAS



SAAS learns on the properties present in the reviewed UniProtKB (Swiss-Prot) entries and uses the following attribute types to define the learning entries: InterPro protein family, taxonomy and sequence length. This combination allows SAAS to generate rules to annotate protein properties such as **function, catalytic activity, pathway membership, subcellular location, protein names and feature predictions.**



SAAS: SAAS00002149

If a protein meets these conditions...

Common conditions

taxon = Bacteria
Matches InterPro signature IPR002146
Matches InterPro signature IPR028987
Matches Pfam signature PF00430
Matches SCOP Superfamily signature SSF81573

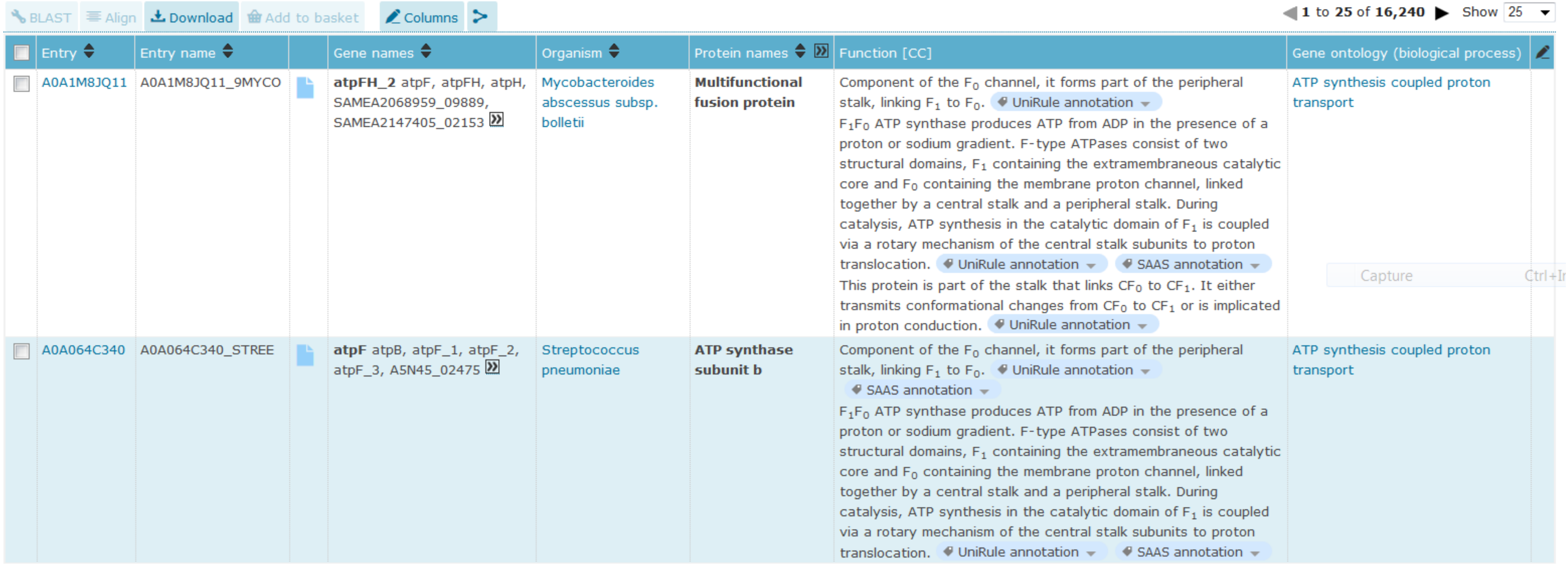
... then these annotations are applied

Function

F_1F_0 ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F_1 containing the extramembraneous catalytic core and F_0 containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F_1 is coupled via a rotary mechanism of the central stalk subunits to proton translocation.

http://insideuniprot.blogspot.ch/2016/10/automatic-learning-based-annotation-in.html

SAAS



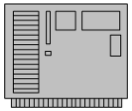
□	Entry	Entry name		Gene names	Organism	Protein names	Function [CC]	Gene ontology (biological process)
□	A0A1M8JQ11	A0A1M8JQ11_9MYCO		**atpFH_2** atpF, atpFH, atpH, SAMEA2068959_09889, SAMEA2147405_02153	Mycobacteroides abscessus subsp. bolletii	**Multifunctional fusion protein**	Component of the F_0 channel, it forms part of the peripheral stalk, linking F_1 to F_0. UniRule annotation F_1F_0 ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F_1 containing the extramembraneous catalytic core and F_0 containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F_1 is coupled via a rotary mechanism of the central stalk subunits to proton translocation. UniRule annotation SAAS annotation This protein is part of the stalk that links CF_0 to CF_1. It either transmits conformational changes from CF_0 to CF_1 or is implicated in proton conduction. UniRule annotation	ATP synthesis coupled proton transport
□	A0A064C340	A0A064C340_STREE		**atpF** atpB, atpF_1, atpF_2, atpF_3, A5N45_02475	Streptococcus pneumoniae	**ATP synthase subunit b**	Component of the F_0 channel, it forms part of the peripheral stalk, linking F_1 to F_0. UniRule annotation SAAS annotation F_1F_0 ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F_1 containing the extramembraneous catalytic core and F_0 containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F_1 is coupled via a rotary mechanism of the central stalk subunits to proton translocation. UniRule annotation SAAS annotation	ATP synthesis coupled proton transport

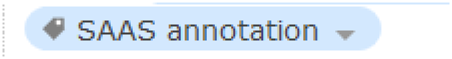
Query: source: SAAS00002149, November 2018

Beware: the SAAS number may change, if the rules changes…

Automated annotation

Automated generated rules (SAAS)

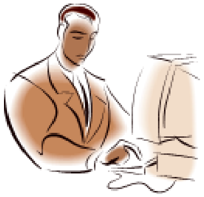




SAAS annotation

Generates a set of decision trees using data mining (new set every UniProtKB release)

Manually generated rules (UniRule)

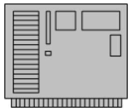


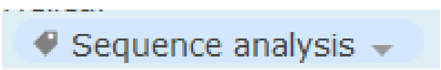
UniRule annotation

Maintains a set of manual annotation rules

UniRule = PIR + HAMAP + Rulebase

Sequence analysis methods (SAM)

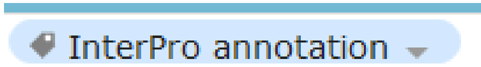




Sequence analysis

Signal, transmembrane, coils prediction

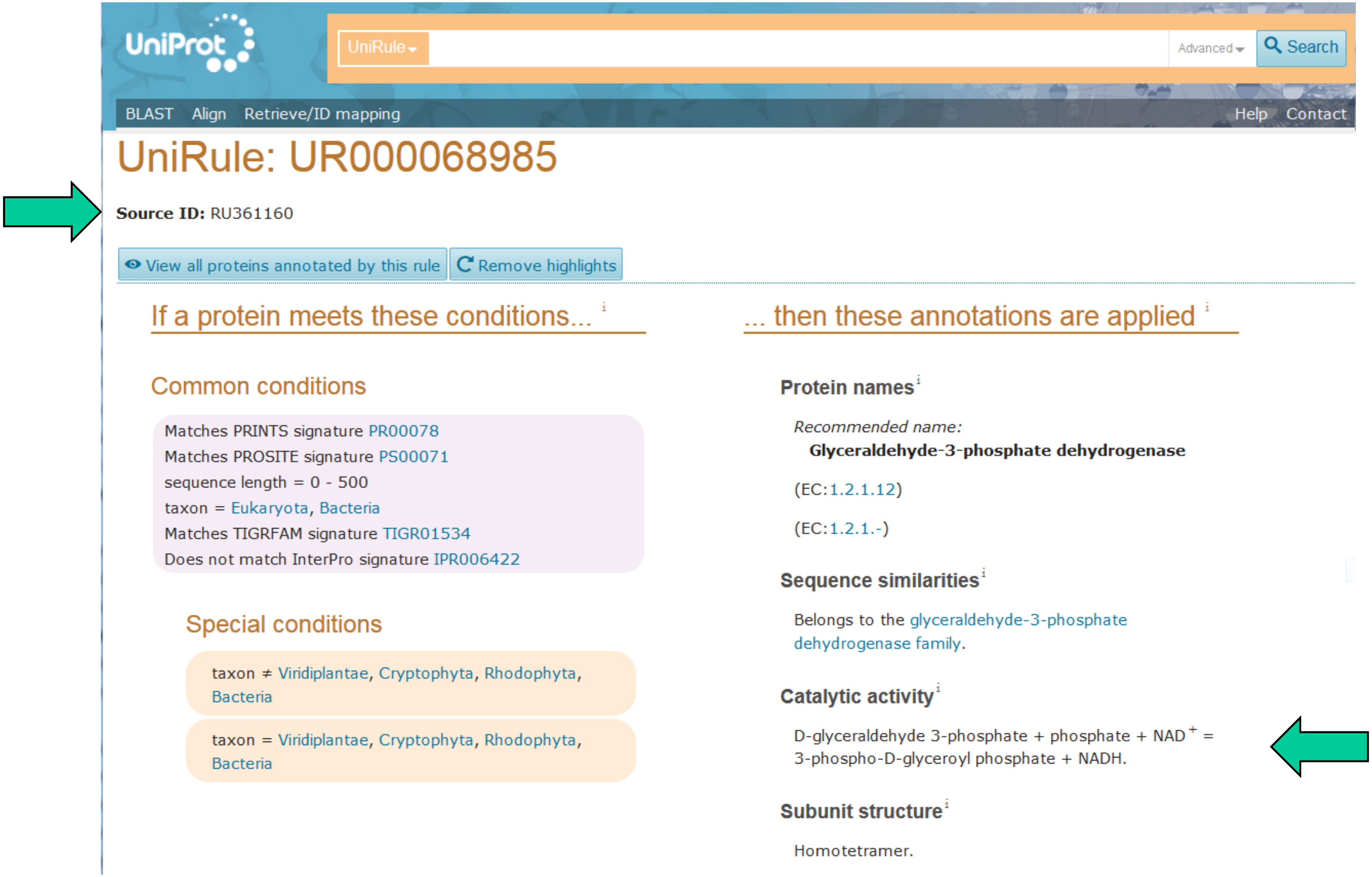
InterPro



InterPro annotation

Domains & GO terms

UniRule (1)



UniRule: UR000068985

Source ID: RU361160

If a protein meets these conditions...

Common conditions

Matches PRINTS signature PR00078
Matches PROSITE signature PS00071
sequence length = 0 - 500
taxon = Eukaryota, Bacteria
Matches TIGRFAM signature TIGR01534
Does not match InterPro signature IPR006422

Special conditions

taxon ≠ Viridiplantae, Cryptophyta, Rhodophyta, Bacteria

taxon = Viridiplantae, Cryptophyta, Rhodophyta, Bacteria

... then these annotations are applied

Protein names

Recommended name:
Glyceraldehyde-3-phosphate dehydrogenase

(EC:1.2.1.12)

(EC:1.2.1.-)

Sequence similarities

Belongs to the glyceraldehyde-3-phosphate dehydrogenase family.

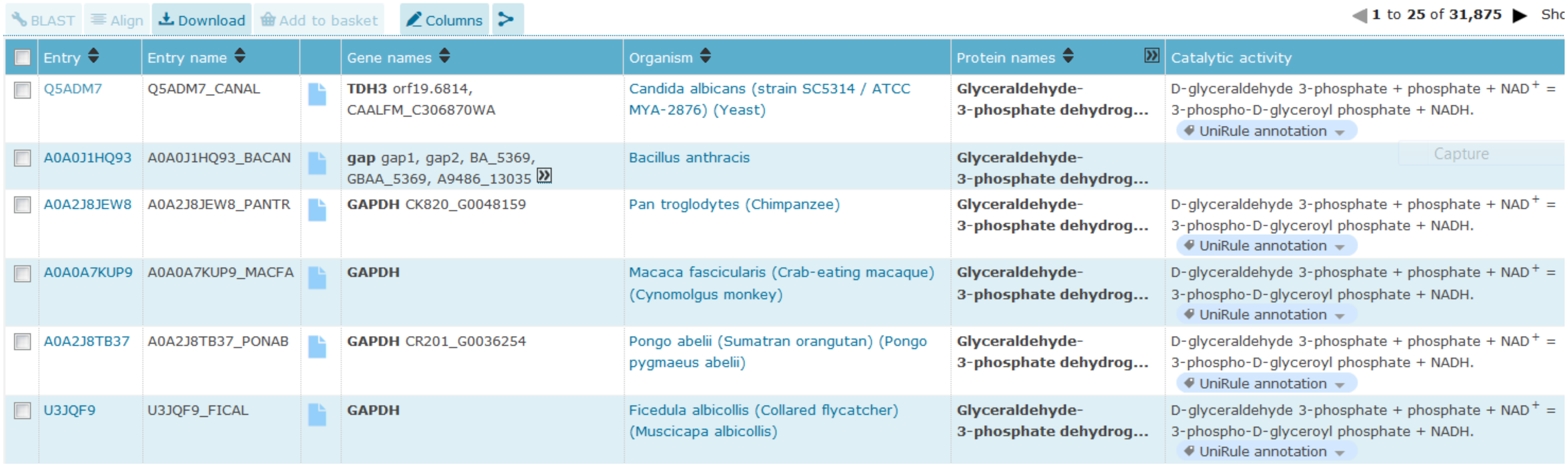
Catalytic activity

D-glyceraldehyde 3-phosphate + phosphate + NAD^+ = 3-phospho-D-glyceroyl phosphate + NADH.

Subunit structure

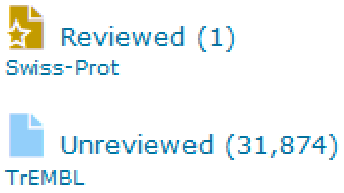
Homotetramer.

UniRule (1)



□	Entry	Entry name		Gene names	Organism	Protein names	Catalytic activity
□	Q5ADM7	Q5ADM7_CANAL		**TDH3** orf19.6814, CAALFM_C306870WA	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)	**Glyceraldehyde-3-phosphate dehydrog...**	D-glyceraldehyde 3-phosphate + phosphate + NAD$^+$ = 3-phospho-D-glyceroyl phosphate + NADH. UniRule annotation
□	A0A0J1HQ93	A0A0J1HQ93_BACAN		**gap** gap1, gap2, BA_5369, GBAA_5369, A9486_13035	Bacillus anthracis	**Glyceraldehyde-3-phosphate dehydrog...**	Capture
□	A0A2J8JEW8	A0A2J8JEW8_PANTR		**GAPDH** CK820_G0048159	Pan troglodytes (Chimpanzee)	**Glyceraldehyde-3-phosphate dehydrog...**	D-glyceraldehyde 3-phosphate + phosphate + NAD$^+$ = 3-phospho-D-glyceroyl phosphate + NADH. UniRule annotation
□	A0A0A7KUP9	A0A0A7KUP9_MACFA		**GAPDH**	Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)	**Glyceraldehyde-3-phosphate dehydrog...**	D-glyceraldehyde 3-phosphate + phosphate + NAD$^+$ = 3-phospho-D-glyceroyl phosphate + NADH. UniRule annotation
□	A0A2J8TB37	A0A2J8TB37_PONAB		**GAPDH** CR201_G0036254	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)	**Glyceraldehyde-3-phosphate dehydrog...**	D-glyceraldehyde 3-phosphate + phosphate + NAD$^+$ = 3-phospho-D-glyceroyl phosphate + NADH. UniRule annotation
□	U3JQF9	U3JQF9_FICAL		**GAPDH**	Ficedula albicollis (Collared flycatcher) (Muscicapa albicollis)	**Glyceraldehyde-3-phosphate dehydrog...**	D-glyceraldehyde 3-phosphate + phosphate + NAD$^+$ = 3-phospho-D-glyceroyl phosphate + NADH. UniRule annotation

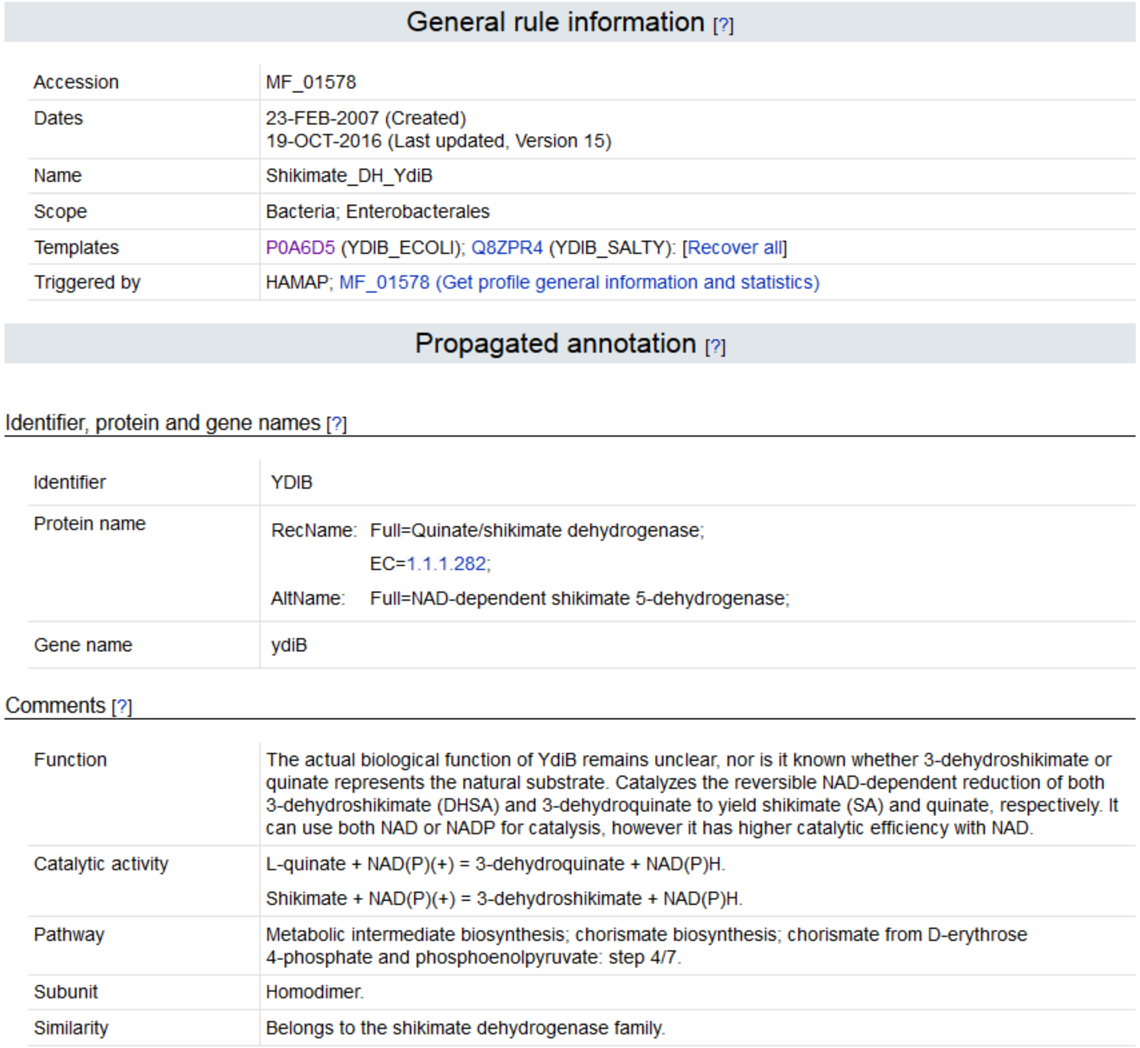
Query: source:RU361160, Nov 2018


Reviewed (1)
Swiss-Prot

Unreviewed (31,874)
TrEMBL

UniRule (2)

Profiles and annotation rules are manually curated: HAMAP



General rule information [?]

Accession	MF_01578
Dates	23-FEB-2007 (Created) 19-OCT-2016 (Last updated, Version 15)
Name	Shikimate_DH_YdiB
Scope	Bacteria; Enterobacterales
Templates	P0A6D5 (YDIB_ECOLI); Q8ZPR4 (YDIB_SALTY); [Recover all]
Triggered by	HAMAP; MF_01578 (Get profile general information and statistics)

Propagated annotation [?]

Identifier, protein and gene names [?]

Identifier	YDIB

Protein name	RecName: Full=Quinate/shikimate dehydrogenase;
	EC=1.1.1.282;
	AltName: Full=NAD-dependent shikimate 5-dehydrogenase;

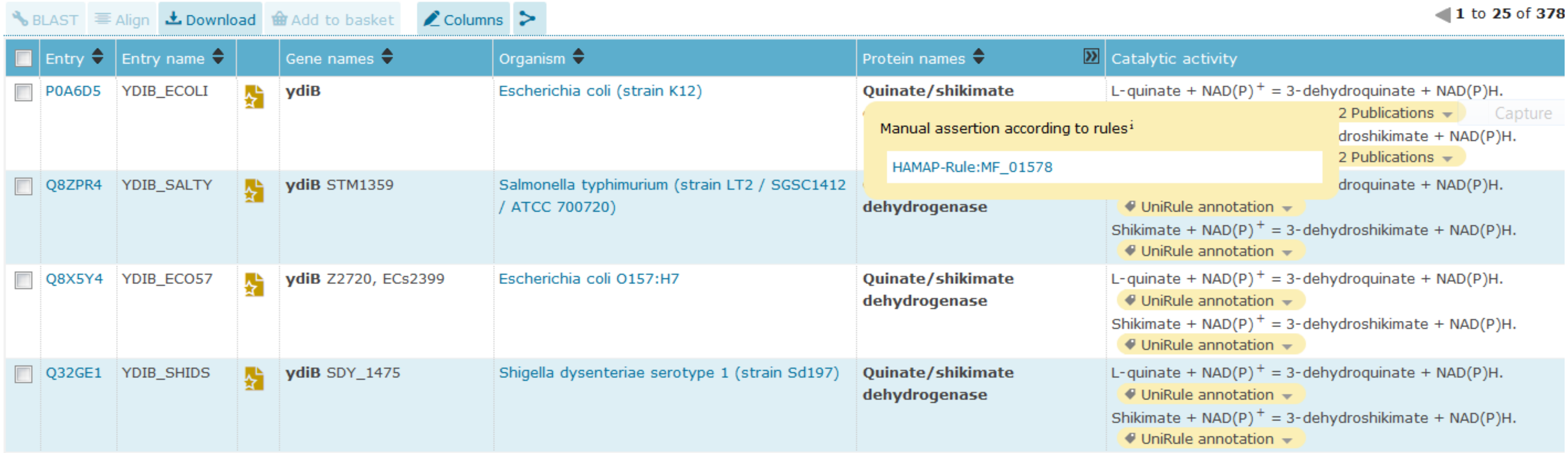
Gene name	ydiB

Comments [?]

Function	The actual biological function of YdiB remains unclear, nor is it known whether 3-dehydroshikimate or quinate represents the natural substrate. Catalyzes the reversible NAD-dependent reduction of both 3-dehydroshikimate (DHSA) and 3-dehydroquinate to yield shikimate (SA) and quinate, respectively. It can use both NAD or NADP for catalysis, however it has higher catalytic efficiency with NAD.
Catalytic activity	L-quinate + NAD(P)(+) = 3-dehydroquinate + NAD(P)H. Shikimate + NAD(P)(+) = 3-dehydroshikimate + NAD(P)H.
Pathway	Metabolic intermediate biosynthesis; chorismate biosynthesis; chorismate from D-erythrose 4-phosphate and phosphoenolpyruvate: step 4/7.
Subunit	Homodimer.
Similarity	Belongs to the shikimate dehydrogenase family.

UniRule (2)

Profiles and annotation rules are manually curated: **HAMAP**

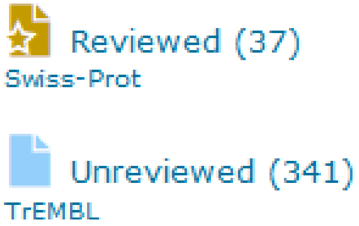


☐	Entry	Entry name		Gene names	Organism	Protein names	Catalytic activity
☐	P0A6D5	YDIB_ECOLI	★	**ydiB**	Escherichia coli (strain K12)	**Quinate/shikimate**	L-quinate + NAD(P)$^+$ = 3-dehydroquinate + NAD(P)H. 2 Publications droshikimate + NAD(P)H. 2 Publications
☐	Q8ZPR4	YDIB_SALTY	★	**ydiB** STM1359	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	**dehydrogenase**	droquinate + NAD(P)H. UniRule annotation Shikimate + NAD(P)$^+$ = 3-dehydroshikimate + NAD(P)H. UniRule annotation
☐	Q8X5Y4	YDIB_ECO57	★	**ydiB** Z2720, ECs2399	Escherichia coli O157:H7	**Quinate/shikimate dehydrogenase**	L-quinate + NAD(P)$^+$ = 3-dehydroquinate + NAD(P)H. UniRule annotation Shikimate + NAD(P)$^+$ = 3-dehydroshikimate + NAD(P)H. UniRule annotation
☐	Q32GE1	YDIB_SHIDS	★	**ydiB** SDY_1475	Shigella dysenteriae serotype 1 (strain Sd197)	**Quinate/shikimate dehydrogenase**	L-quinate + NAD(P)$^+$ = 3-dehydroquinate + NAD(P)H. UniRule annotation Shikimate + NAD(P)$^+$ = 3-dehydroshikimate + NAD(P)H. UniRule annotation

Manual assertion according to rules

HAMAP-Rule:MF_01578

Query: source:"mf 01578"

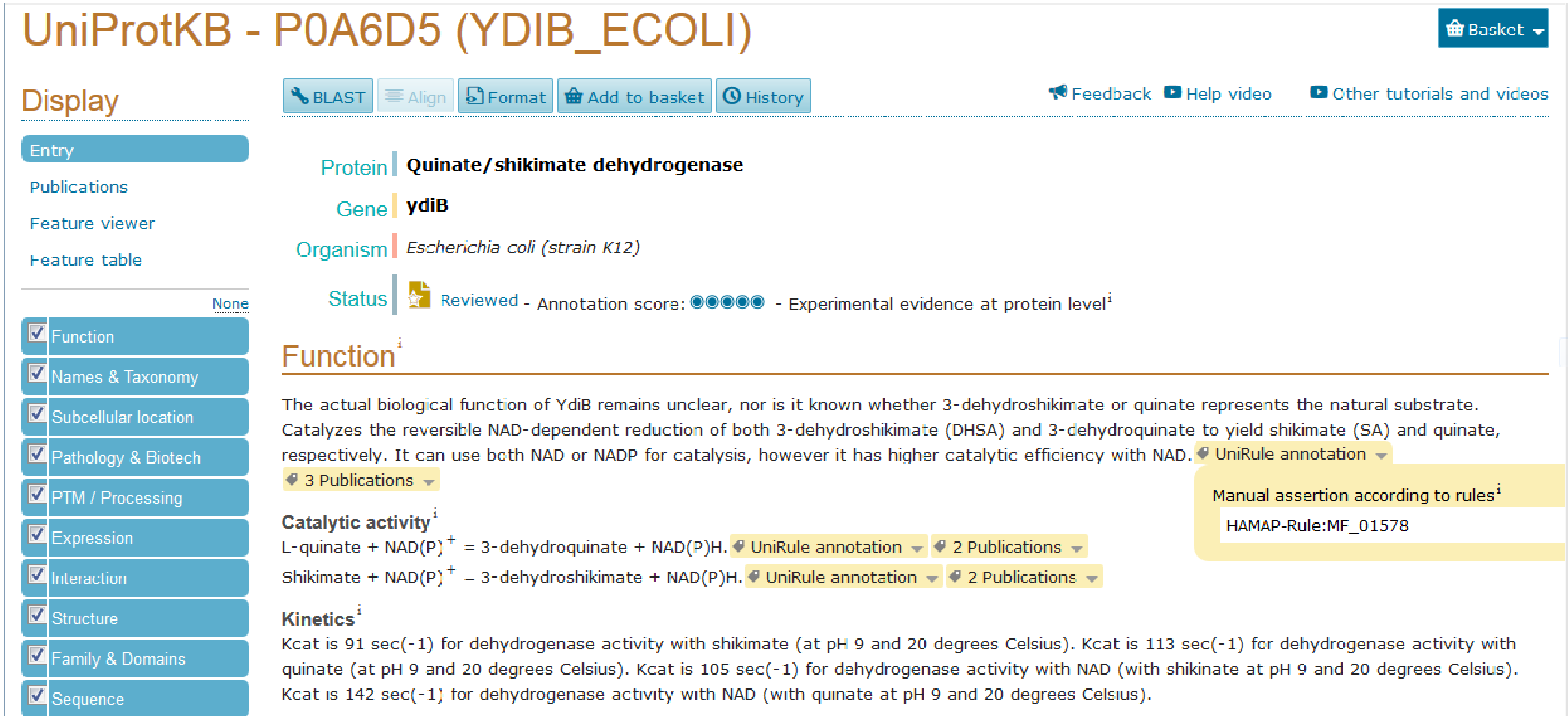


Reviewed (37)
Swiss-Prot

Unreviewed (341)
TrEMBL

UniRule (2)

Profiles and annotation rules are manually curated: HAMAP



UniProtKB - P0A6D5 (YDIB_ECOLI)

Protein | Quinate/shikimate dehydrogenase

Gene | ydiB

Organism | *Escherichia coli (strain K12)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Function[i]

The actual biological function of YdiB remains unclear, nor is it known whether 3-dehydroshikimate or quinate represents the natural substrate. Catalyzes the reversible NAD-dependent reduction of both 3-dehydroshikimate (DHSA) and 3-dehydroquinate to yield shikimate (SA) and quinate, respectively. It can use both NAD or NADP for catalysis, however it has higher catalytic efficiency with NAD. UniRule annotation 3 Publications

Manual assertion according to rules[i]

HAMAP-Rule:MF_01578

Catalytic activity[i]

L-quinate + NAD(P)$^+$ = 3-dehydroquinate + NAD(P)H. UniRule annotation 2 Publications

Shikimate + NAD(P)$^+$ = 3-dehydroshikimate + NAD(P)H. UniRule annotation 2 Publications

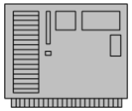
Kinetics[i]

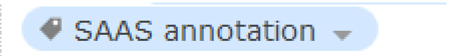
Kcat is 91 sec(-1) for dehydrogenase activity with shikimate (at pH 9 and 20 degrees Celsius). Kcat is 113 sec(-1) for dehydrogenase activity with quinate (at pH 9 and 20 degrees Celsius). Kcat is 105 sec(-1) for dehydrogenase activity with NAD (with shikinate at pH 9 and 20 degrees Celsius). Kcat is 142 sec(-1) for dehydrogenase activity with NAD (with quinate at pH 9 and 20 degrees Celsius).

Swiss Institute of Bioinformatics

Automated annotation

Automated generated rules (SAAS)

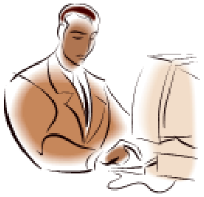




Generates a set of decision trees using data mining (new set every UniProtKB release)

Manually generated rules (UniRule)

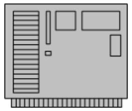
Maintains a set of manual annotation rules

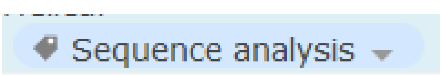


UniRule = PIR + HAMAP + Rulebase

Sequence analysis methods (SAM)

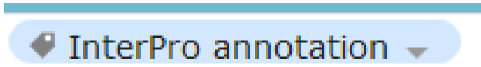
Signal, transmembrane, coils prediction



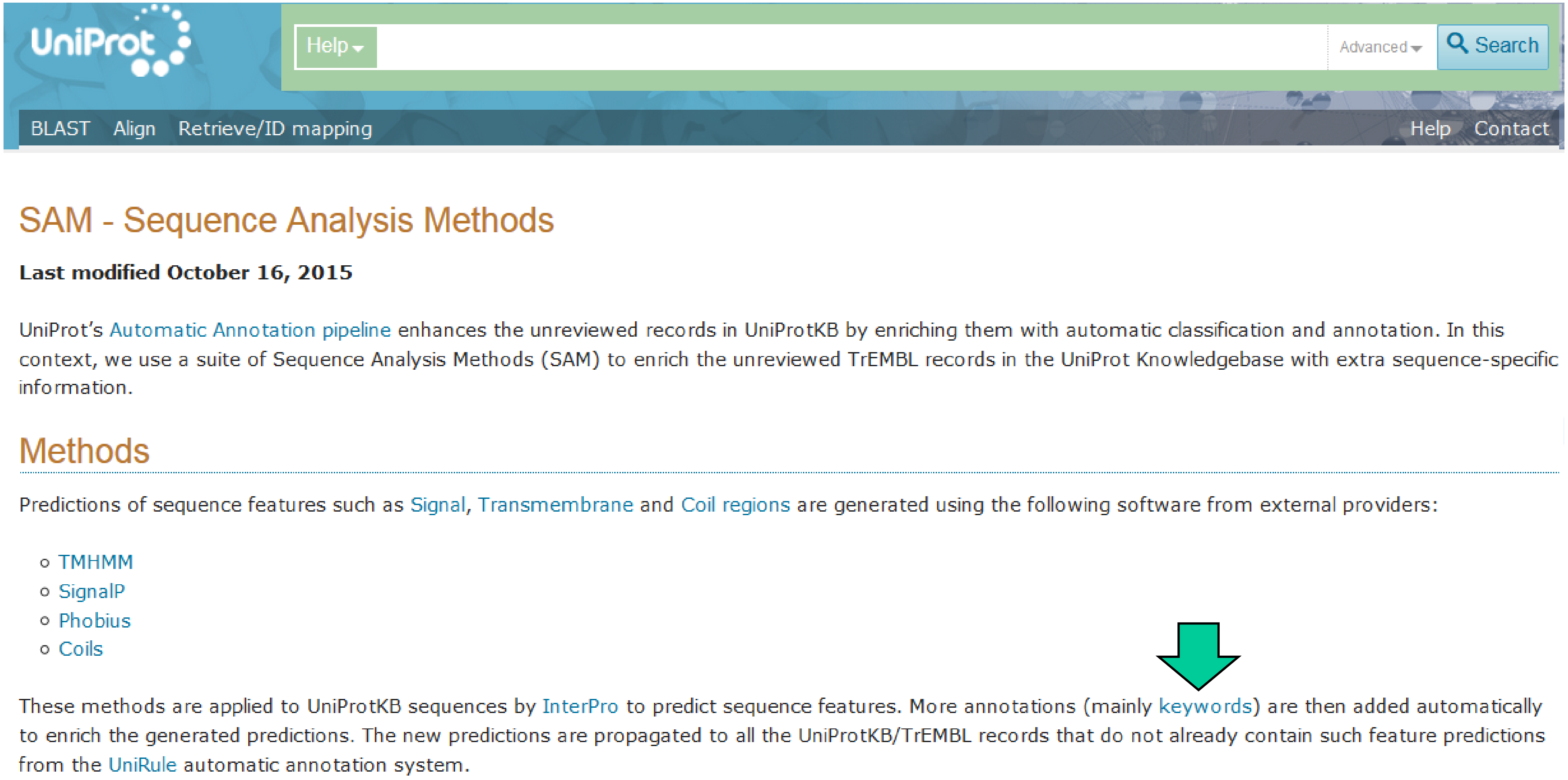


InterPro

Domains & GO terms



SAM



SAM - Sequence Analysis Methods

Last modified October 16, 2015

UniProt's Automatic Annotation pipeline enhances the unreviewed records in UniProtKB by enriching them with automatic classification and annotation. In this context, we use a suite of Sequence Analysis Methods (SAM) to enrich the unreviewed TrEMBL records in the UniProt Knowledgebase with extra sequence-specific information.

Methods

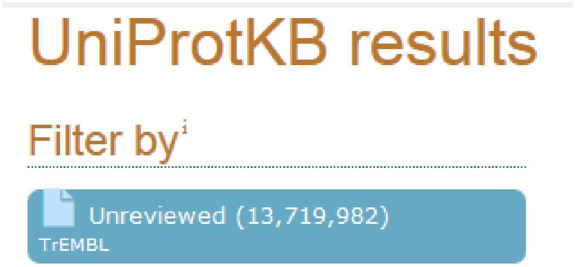
Predictions of sequence features such as Signal, Transmembrane and Coil regions are generated using the following software from external providers:

- TMHMM
- SignalP
- Phobius
- Coils

These methods are applied to UniProtKB sequences by InterPro to predict sequence features. More annotations (mainly keywords) are then added automatically to enrich the generated predictions. The new predictions are propagated to all the UniProtKB/TrEMBL records that do not already contain such feature predictions from the UniRule automatic annotation system.

No GO (cellular component) annotation in this case…
http://www.uniprot.org/help/sam

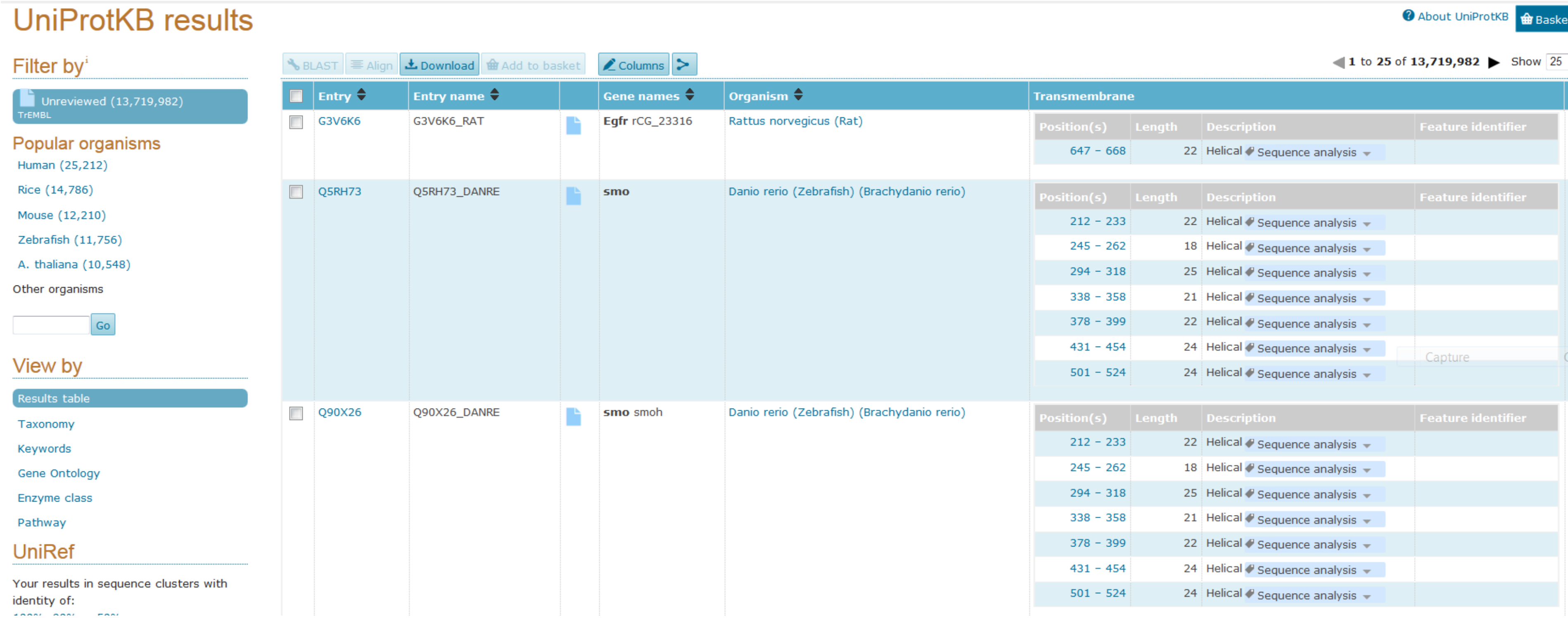
SAM: Transmembrane



UniProtKB results

Filter by

Unreviewed (13,719,982)
TrEMBL



UniProtKB results

1 to 25 of 13,719,982 ▶ Show 25

Filter by

Unreviewed (13,719,982)
TrEMBL

Popular organisms

- Human (25,212)
- Rice (14,786)
- Mouse (12,210)
- Zebrafish (11,756)
- A. thaliana (10,548)

Other organisms

View by

- Results table
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

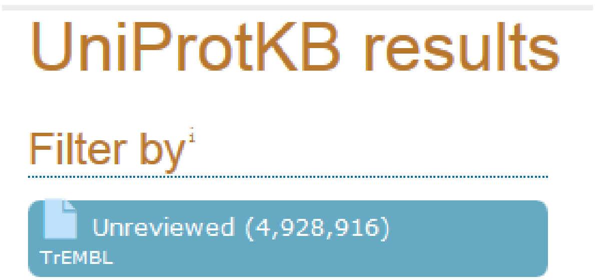
Your results in sequence clusters with identity of:

☐	Entry	Entry name	Gene names	Organism	Transmembrane Position(s)	Length	Description	Feature identifier
☐	G3V6K6	G3V6K6_RAT	**Egfr** rCG_23316	Rattus norvegicus (Rat)	647 – 668	22	Helical Sequence analysis	
☐	Q5RH73	Q5RH73_DANRE	**smo**	Danio rerio (Zebrafish) (Brachydanio rerio)	212 – 233	22	Helical Sequence analysis	
					245 – 262	18	Helical Sequence analysis	
					294 – 318	25	Helical Sequence analysis	
					338 – 358	21	Helical Sequence analysis	
					378 – 399	22	Helical Sequence analysis	
					431 – 454	24	Helical Sequence analysis	Capture
					501 – 524	24	Helical Sequence analysis	
☐	Q90X26	Q90X26_DANRE	**smo** smoh	Danio rerio (Zebrafish) (Brachydanio rerio)	212 – 233	22	Helical Sequence analysis	
					245 – 262	18	Helical Sequence analysis	
					294 – 318	25	Helical Sequence analysis	
					338 – 358	21	Helical Sequence analysis	
					378 – 399	22	Helical Sequence analysis	
					431 – 454	24	Helical Sequence analysis	
					501 – 524	24	Helical Sequence analysis	

annotation:(type:transmem) AND reviewed:no

Swiss Institute of Bioinformatics

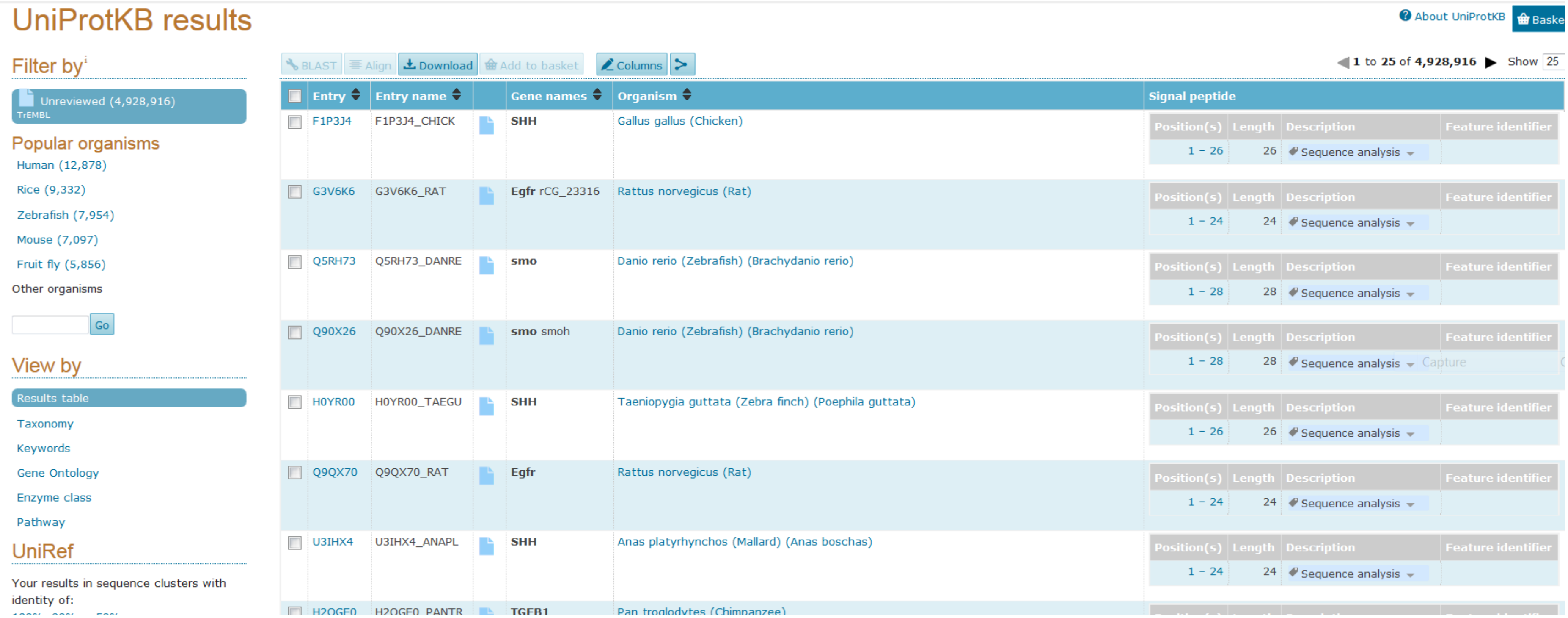
SAM: Signal peptide



UniProtKB results

Filter by

Unreviewed (4,928,916)
TrEMBL



UniProtKB results

About UniProtKB | Baske

Filter by

Unreviewed (4,928,916)
TrEMBL

Popular organisms

- Human (12,878)
- Rice (9,332)
- Zebrafish (7,954)
- Mouse (7,097)
- Fruit fly (5,856)

Other organisms

Go

View by

- Results table
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

Your results in sequence clusters with identity of:

BLAST | Align | Download | Add to basket | Columns

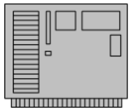
1 to 25 of 4,928,916 ▶ Show 25

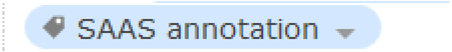
□	Entry	Entry name		Gene names	Organism	Signal peptide			
						Position(s)	Length	Description	Feature identifier
□	F1P3J4	F1P3J4_CHICK		**SHH**	Gallus gallus (Chicken)	1 – 26	26	Sequence analysis	
□	G3V6K6	G3V6K6_RAT		**Egfr** rCG_23316	Rattus norvegicus (Rat)	1 – 24	24	Sequence analysis	
□	Q5RH73	Q5RH73_DANRE		**smo**	Danio rerio (Zebrafish) (Brachydanio rerio)	1 – 28	28	Sequence analysis	
□	Q90X26	Q90X26_DANRE		**smo** smoh	Danio rerio (Zebrafish) (Brachydanio rerio)	1 – 28	28	Sequence analysis	Capture
□	H0YR00	H0YR00_TAEGU		**SHH**	Taeniopygia guttata (Zebra finch) (Poephila guttata)	1 – 26	26	Sequence analysis	
□	Q9QX70	Q9QX70_RAT		**Egfr**	Rattus norvegicus (Rat)	1 – 24	24	Sequence analysis	
□	U3IHX4	U3IHX4_ANAPL		**SHH**	Anas platyrhynchos (Mallard) (Anas boschas)	1 – 24	24	Sequence analysis	
□	H2QGF0	H2QGF0_PANTR		**TGFB1**	Pan troglodytes (Chimpanzee)				

Swiss Institute of Bioinformatics

Automated annotation

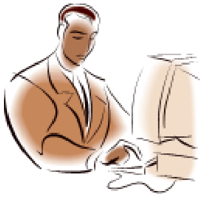
Automated generated rules (SAAS)




SAAS annotation

Generates a set of decision trees using data mining (new set every UniProtKB release)

Manually generated rules (UniRule)

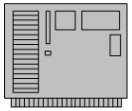


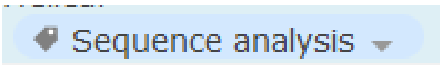
UniRule annotation

Maintains a set of manual annotation rules

UniRule = PIR* + HAMAP + Rulebase

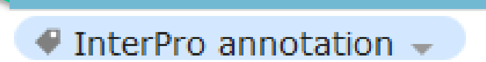
Sequence analysis methods (SAM)



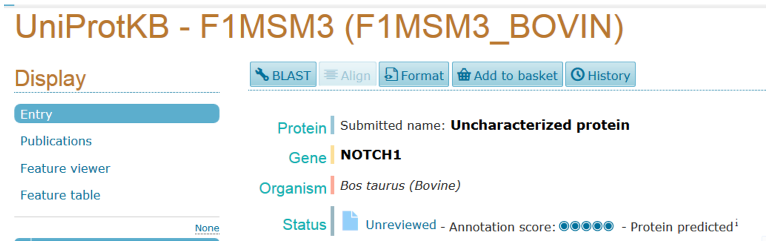

Sequence analysis

Signal, transmembrane, coils prediction

InterPro


InterPro annotation

Domains & GO terms



UniProtKB - F1MSM3 (F1MSM3_BOVIN)

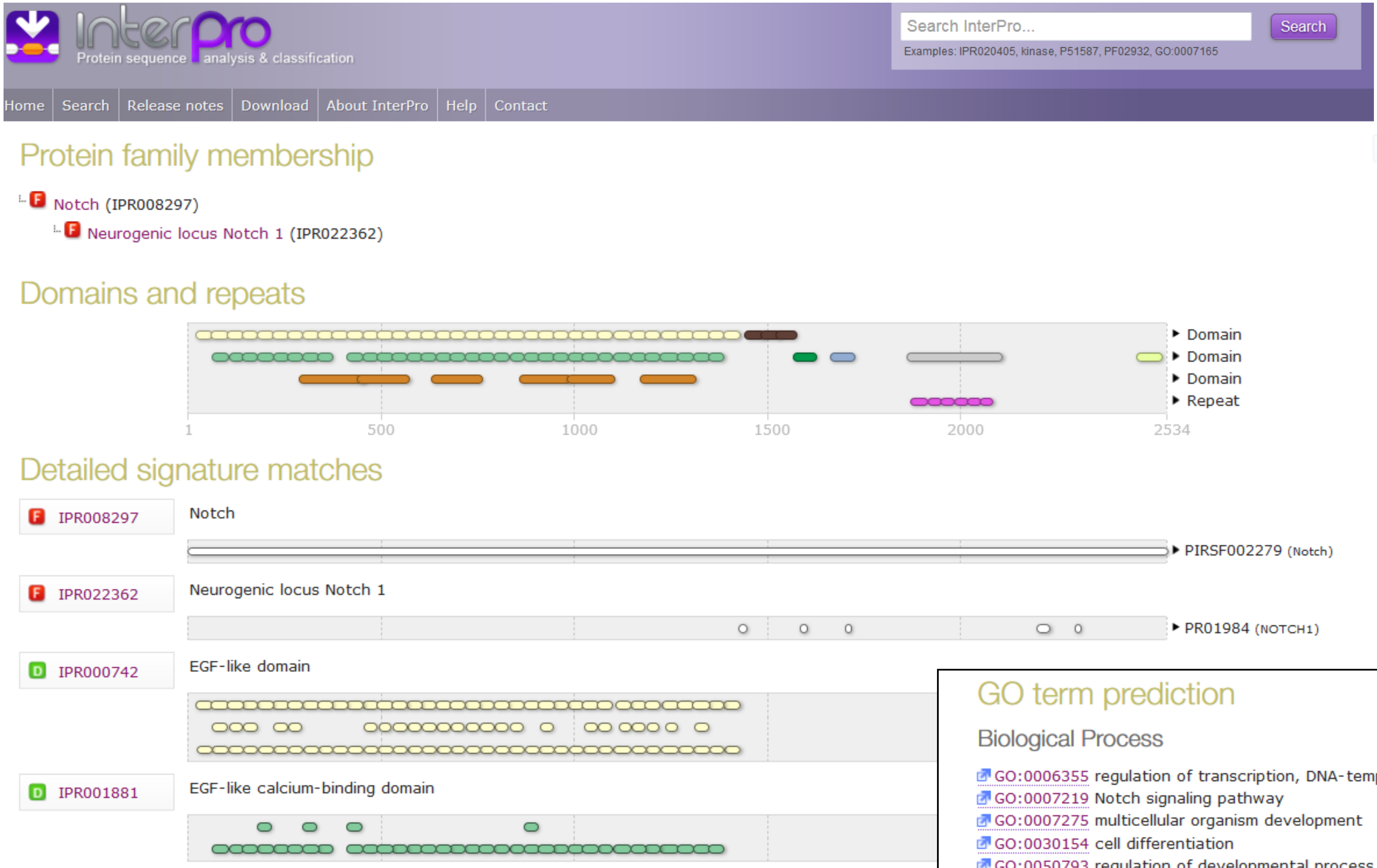
Protein | Submitted name: **Uncharacterized protein**

Gene | **NOTCH1**

Organism | *Bos taurus (Bovine)*

Status | Unreviewed - Annotation score: ●●●●● - Protein predicted[i]

```
>tr|F1MSM3|F1MSM3_BOVIN Uncharacterized protein OS=Bos taurus GN=NOTCH1 PE=4 SV=2
MPPLLAPLLCLALLPALAARGLRCSQPGETCLNGGKCEVFPNGTEACICGGAFAGQQCQA
PNPCLSAPCKNGGTCHTTEREGLVDYVCGCRLGFSGPLCLTPRDHACLASPCLNGGTCDL
LTLTEYKCLCTPGWSGKTCQQADPCASNPCANGGQCLPFEASYICHCPPGFHGPTCRQDV
NECSQSPGLCHHGGTCLNEVGSYRCVCRPTHTGPHCELPYVPCSPSPCQNGGTCRPTGDT
THECACLPGFTGQNCEENIDDCPGNSCKNGGACVDGVNTYNCRCPPEWTGQYCTEDVDEC
QLMPNACQNGGTCHNTHGGYNCVCVNGWTGEDCSENIDDCASASCFQGATCHDRVASFYC
ECPHGRTGLLCHLNDACISNPCNEGSNCDTNPVNGKAICTCPSGYTGPACSQDVDECSLG
ANPCEHAGKCINTLGSFECQCLQGYTGPRCEIDVNECVSNPCQNDATCLDQIGEFQCICM
PGYEGLHCEVNTDECASSPCLQNGRCLDKINEFVCECPTGFTGHLCQYDVDECASTPCKN
GAKCLDGPNTYTCVCTEGYTGPHCEVDIDECDPDPCHYGSCKDGVATFTCLCQPGYTGHH
CESNINECHSQPCRHGGTCQDRDNAYLCFCLKGTTGPNCEINLDDCASNPCDSGTCLDKI
DGYECACEPGYTGSMCNINIDECADSPCHNGGTCEDGINGFTCRCPEGYHDPTCLSEVNE
CSSNPCIHGACRDSLNGYKCDCDPGWSGANCDVNNDECESNPCINGGTCKDMTSGYVCAC
REGFSGPNCQTNINECASNPCLNQGTCIDDVAGYKCNCLLPYTGATCEVVLAPCAPGPCR
NGGECRESEDYESFSCACPAGWQGQTCEIDINECVKSPCRAGASCQNTNGSYRCHCQAGY
TGRNCETDIDDCRPNPCHNGGSCTDGINTAFCDCLPGFQGAFCEEDINECASSPCRNGAN
CTDCVDSYTCTCPTGFSGIHCENNTPDCTESSCFNGGTCVDGINSFTCLCPPGFTGSYCQ
HDVNECDSRPCLHGGTCHDSYGTYTCTCPQGYTGLNCQTLVRWCDSSPCKNDGRCWQTNA
LYRCECHSGWTGLYCDVPSVSCEVAARQQGVNVTHLCRNGGLCMNAGNTHRCHCQAGYTG
SYCEEQVDECSPSPCQNGATCTDYPGGYSCECVAGYHGVNCSEEVNECLSQPCRNGGTCI
DLTNTYKCSCPRGTQGVHCEINVDDCNPPIDPVSRGPKCFNNGTCVDQVGGYSCSCPPGF
VGERCEGDVNECLSNPCDARGTQNCVQHVNAFHCECRAGHTGRRCESVINGCKDRPCKNG
GSCAVASNTARGFICKCPAGFEGATCENDARSCGSLRCLNGGTCIAGPRSPTCLCLGPFT
GPECQFPASSPCVGGNPCYNQGVCEPTAESPFYRCRCPAKFNGLLCHILDYSFGGGVGLD
IPPPQIEETCELPGCREEAGNKVCSLQCNSHACGWDGGDCSLDFDDPWQNCTQSLQCWKY
FSNGRCDSQCNSAGCLFDGFDCQRAEGQCNPLYDQYCKDHFRDGHCDQGCNSAECEWDGL
DCAEHVPERLAAGTLVLVVLMPPEQLRNRSLHFLRELSRLLHTNVVFKRDASGQQMIFPY
YGQEPHCRQGSAPRSVGVSTTHALLVLDKASPQGHCAPPGLFLSIVYLEIDNRQCVQSSS
QCFQSATDVAAFLGALASLGSLNIPYKIEAVQSETVEPPPPPPLHFMYVAVVAFVLLFFV
GCGVLLSRKRRRQHGQLWFPEGFKVSEASKKKRREPLGEDSVGLKPLKNSSDGALMDDNQ
NEWGDEGLEAKKFRFEEPVVLPDLDDQTDHRQWTQQHLDAADLRVSAMAPTPPQGEADAD
CMDVNVRGPDGFTPLMIASCSGGGLETGNSEEEEDAPAVISDFIYQGASLHNQTDRTGET
ALHLAARYSRSDAAKRLLEASADANIQDNMGRTPLHAAVSADAQGVFQILIRNRATDLDA
RMHDGTTPLILAARLAVEGMLEDLINSHADVNAVDDLGKSALHWAAAVNNVEAAVVLLKN
GANKDMQNNKEETPLFLAAREGSYETAKVLLDHFANRDITDHMDRLPRDIAQERMHHDIV
RLLDEYSLVRSPPLHGATLGGTPTLSPPLCSPNGYLGNLKPPMQGKKARKPSTKGLACGG
KEPKDLKARRKKSQDGKGCLLDSGSVMSPVDSLESPHGYLSDVASPPLLPSPFQPSPSVP
LNHLPGMPETHLGVSHLSVAAKPEMAVLSGGSRLAFEAGPPRLSHLPVASSTSTILGSGG
SGGSGAVNFTVGGAAGLNGQCEWLSRLQNGLVPNQYNPLRGGVTPGTLSTQAAGLQHGTV
GPLHAPALSQVMTYQALPSTRLASQPHLVQPQQNLQMQPPSMPPQPNLQPHLGVSSAASG
HLGRSFLGGELSQADMQPLGPGNLAAHTVLPQDGQVLPTSLPSTLAPPTMAPPMTTAQFL
TPPSQHSYSSSPVDNTPSHQLQVPEHPFLTPSPESPDQWSSSSPHSNISDWSEGISSPPT
SVPSQIAHVPEAFK
```



InterPro
Protein sequence analysis & classification
Search InterPro...
Search
Examples: IPR020405, kinase, P51587, PF02932, GO:0007165
Home
Search
Release notes
Download
About InterPro
Help
Contact

Protein family membership

- F Notch (IPR008297)
  - F Neurogenic locus Notch 1 (IPR022362)

Domains and repeats

Detailed signature matches

F IPR008297 Notch

F IPR022362 Neurogenic locus Notch 1

D IPR000742 EGF-like domain

D IPR001881 EGF-like calcium-binding domain

GO term prediction

Biological Process

- GO:0006355 regulation of transcription, DNA-templated
- GO:0007219 Notch signaling pathway
- GO:0007275 multicellular organism development
- GO:0030154 cell differentiation
- GO:0050793 regulation of developmental process

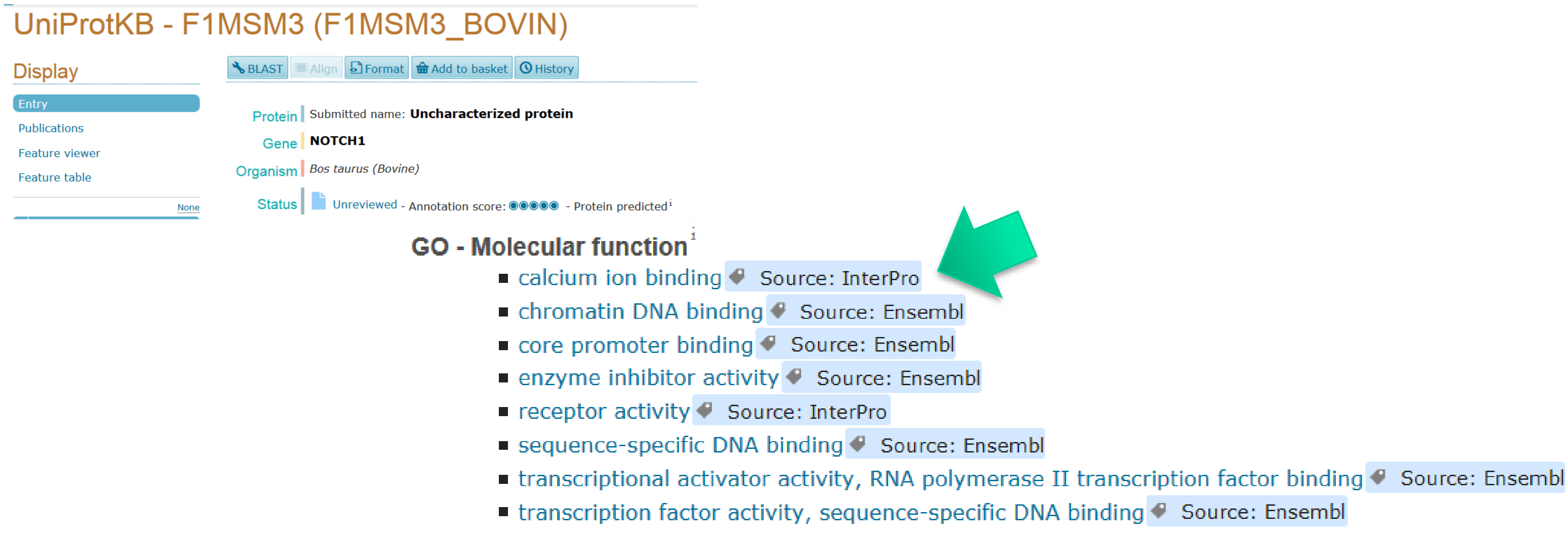
Molecular Function

- GO:0004872 receptor activity
- GO:0005509 calcium ion binding
- GO:0005515 protein binding

Cellular Component

- GO:0016021 integral component of membrane

Prosite, Smart, PFAM



UniProtKB - F1MSM3 (F1MSM3_BOVIN)

Protein | Submitted name: **Uncharacterized protein**

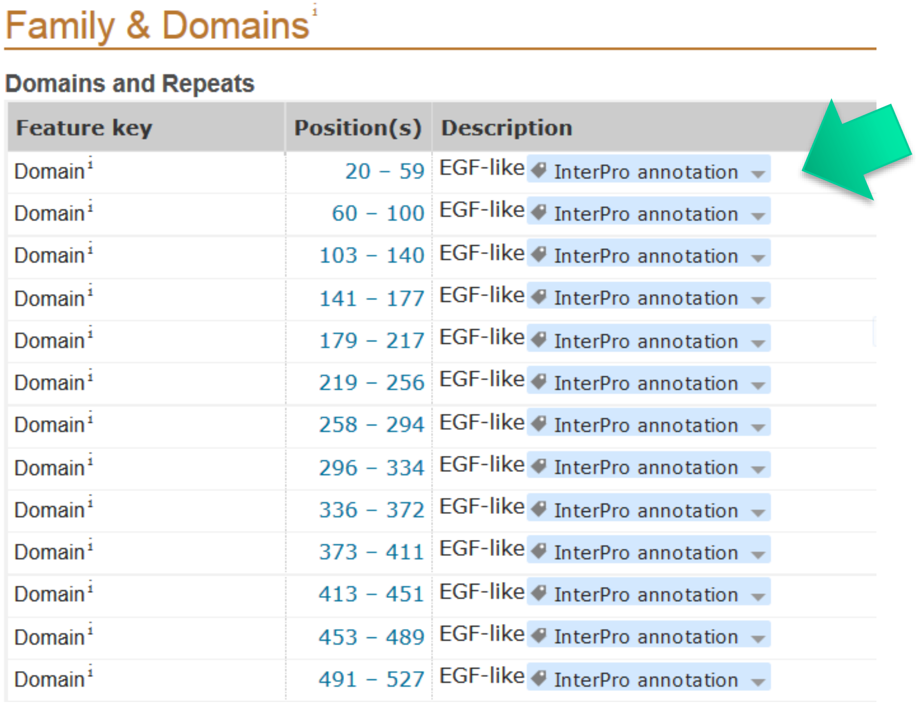
Gene | **NOTCH1**

Organism | *Bos taurus (Bovine)*

Status | Unreviewed - Annotation score: ●●●●● - Protein predicted[i]

GO - Molecular function[i]

- calcium ion binding Source: InterPro
- chromatin DNA binding Source: Ensembl
- core promoter binding Source: Ensembl
- enzyme inhibitor activity Source: Ensembl
- receptor activity Source: InterPro
- sequence-specific DNA binding Source: Ensembl
- transcriptional activator activity, RNA polymerase II transcription factor binding Source: Ensembl
- transcription factor activity, sequence-specific DNA binding Source: Ensembl



Family & Domains[i]

Domains and Repeats

Feature key	Position(s)	Description
Domain[i]	20 – 59	EGF-like InterPro annotation
Domain[i]	60 – 100	EGF-like InterPro annotation
Domain[i]	103 – 140	EGF-like InterPro annotation
Domain[i]	141 – 177	EGF-like InterPro annotation
Domain[i]	179 – 217	EGF-like InterPro annotation
Domain[i]	219 – 256	EGF-like InterPro annotation
Domain[i]	258 – 294	EGF-like InterPro annotation
Domain[i]	296 – 334	EGF-like InterPro annotation
Domain[i]	336 – 372	EGF-like InterPro annotation
Domain[i]	373 – 411	EGF-like InterPro annotation
Domain[i]	413 – 451	EGF-like InterPro annotation
Domain[i]	453 – 489	EGF-like InterPro annotation
Domain[i]	491 – 527	EGF-like InterPro annotation

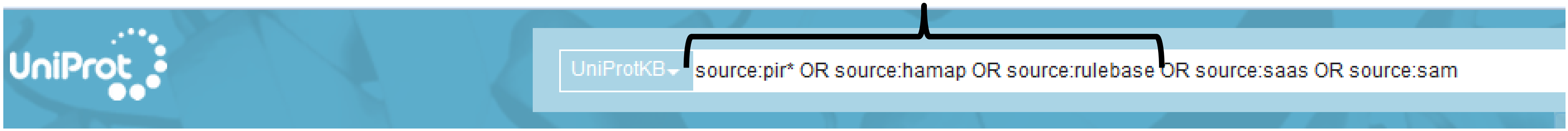
UniProtKB/TrEMBL
&
Automated annotation

Important remarks

UniProt

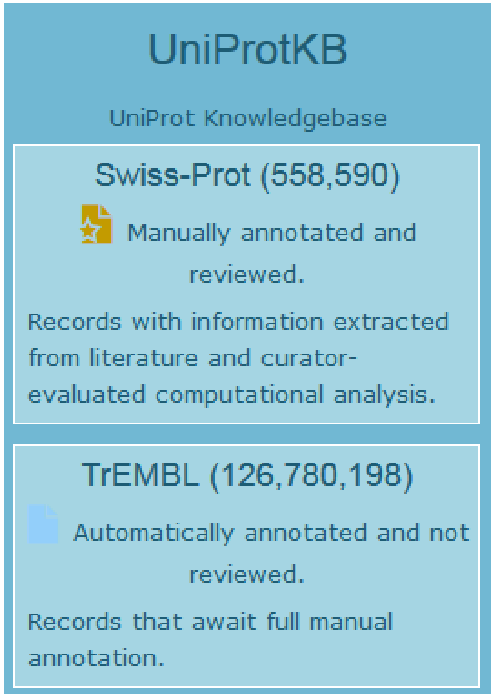
SIB Swiss Institute of Bioinformatics

UniRule



source:pir* OR source:hamap OR source:rulebase OR source:saas OR source:sam

Total number of records in UniProtKB



UniProtKB

UniProt Knowledgebase

Swiss-Prot (558,590)

Manually annotated and reviewed.

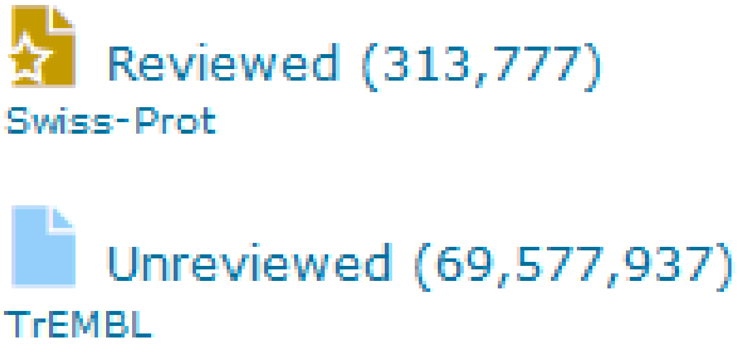
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (126,780,198)

Automatically annotated and not reviewed.

Records that await full manual annotation.

Records with house-made automated annotation



Reviewed (313,777)
Swiss-Prot

Unreviewed (69,577,937)
TrEMBL

Only ~ 55 % of TrEMBL records contain automated annotation

Differences between TrEMBL and Swiss-Prot

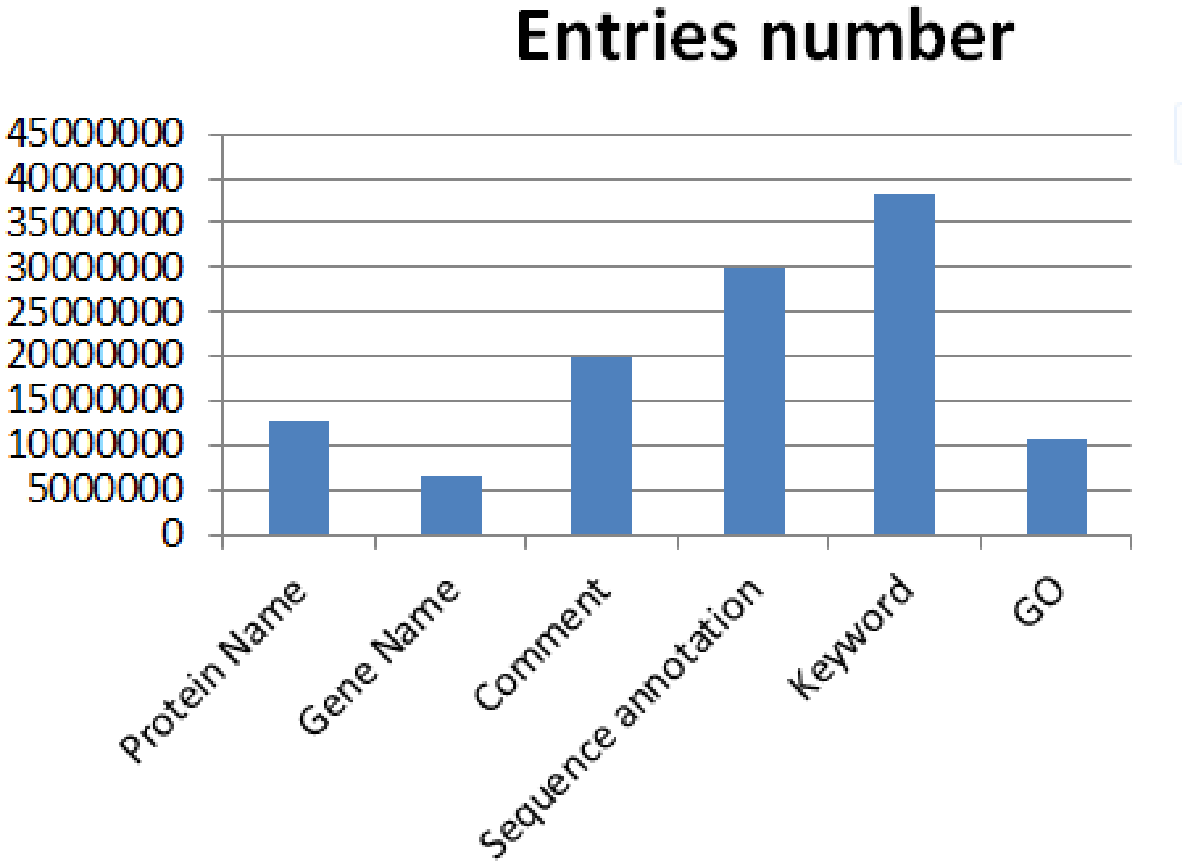
	TrEMBL	Swiss-Prot
annotation	automatic	manual
Annotation = complete ?	Partial annotation (~55 % of the entries)	As complete and systematic as possible
Set of sequences = complete ?	As complete as possible; does not contain Swiss-Prot sequences !	Complete sets only for a few organisms
Number of entries	127 000 000	550 000
Number of species	700 000	13 000

When you compare biological information of given datasets of proteins beware the ratio of TrEMBL vs Swiss-Prot entries in your dataset: the results might not be only 'biological'!

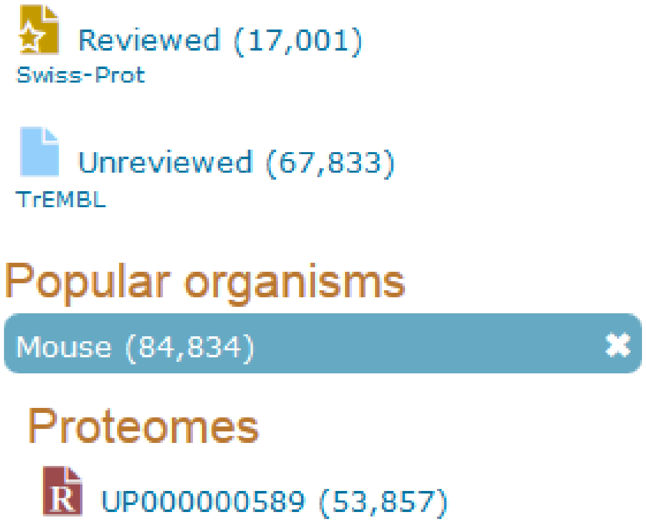
UniProt

SIB Swiss Institute of Bioinformatics

Automated annotation



Set of mouse proteins with N-glycosylation – Nov 2018



Reviewed (17,001)
Swiss-Prot

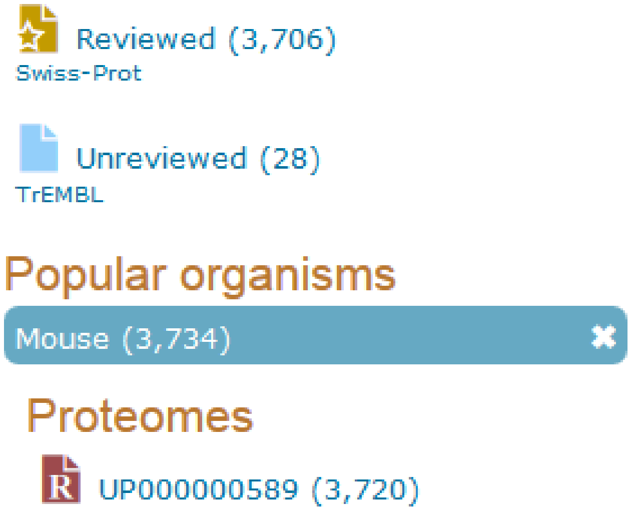
Unreviewed (67,833)
TrEMBL

Popular organisms

Mouse (84,834) ✖

Proteomes

UP000000589 (53,857)



Reviewed (3,706)
Swiss-Prot

Unreviewed (28)
TrEMBL

Popular organisms

Mouse (3,734) ✖

Proteomes

UP000000589 (3,720)

21.7 %

0.04 %

6.0 %

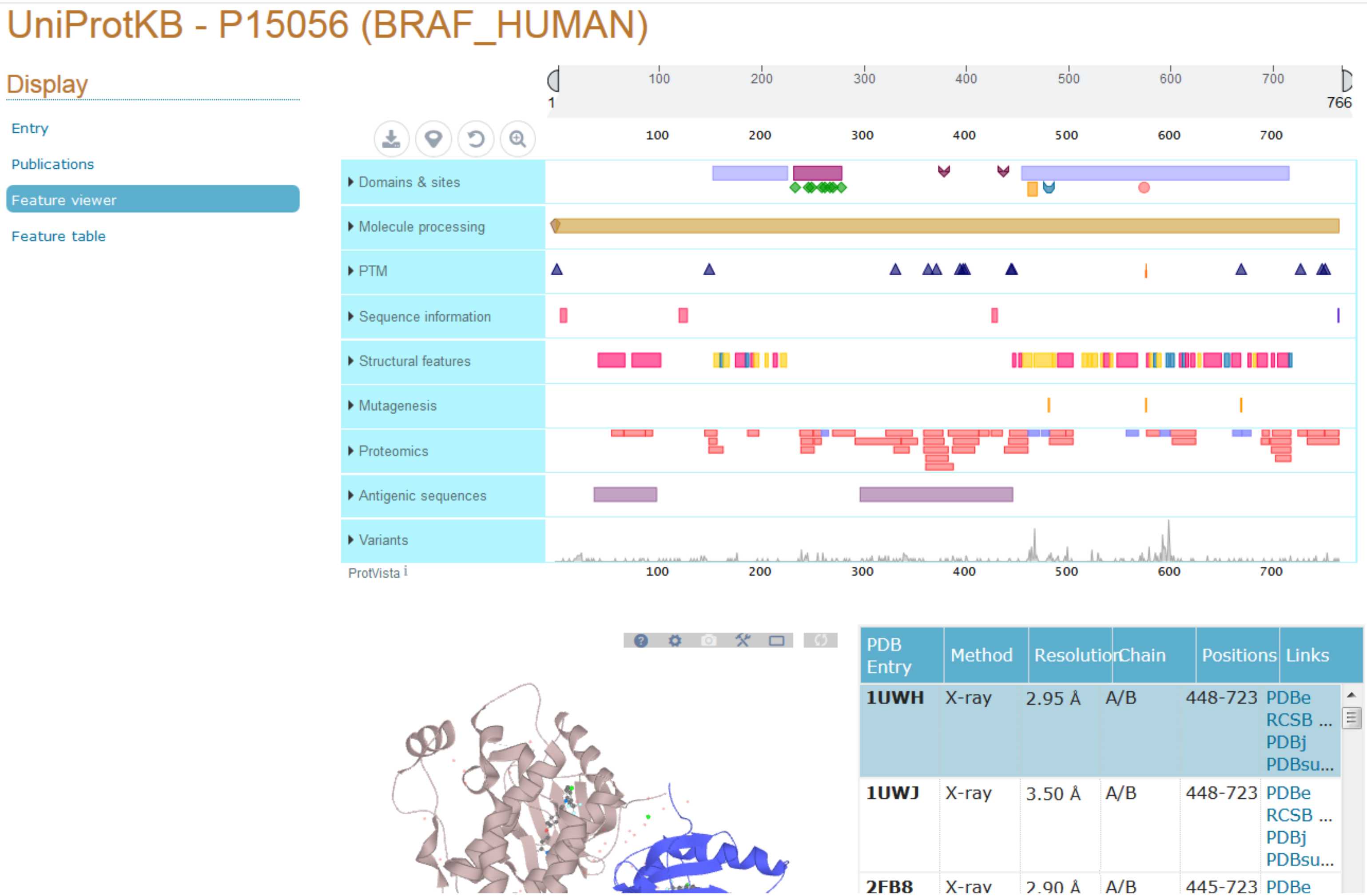
organism:"Mus musculus (Mouse) [10090]"

annotation:(type:carbohyd "n linked glcnac ellipsis") AND organism:"Mus musculus (Mouse) [10090]"

UniProtKB sequence annotation

- Feature viewer
- Genome browser

UniProtKB annotation & Feature viewer



UniProtKB - P15056 (BRAF_HUMAN)

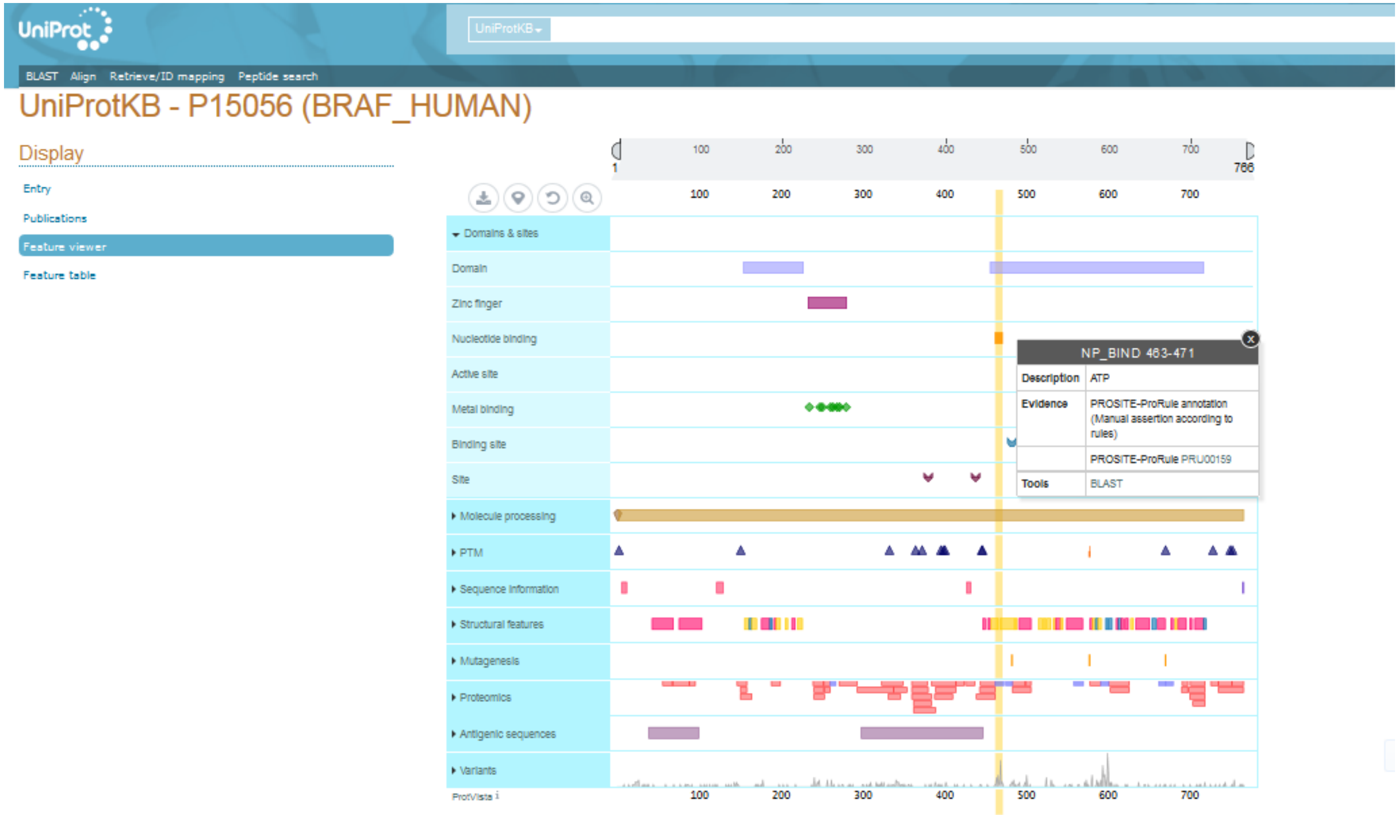
Display

- Entry
- Publications
- Feature viewer
- Feature table

PDB Entry	Method	Resolutio	Chain	Positions	Links
1UWH	X-ray	2.95 Å	A/B	448-723	PDBe RCSB ... PDBj PDBsu...
1UWJ	X-ray	3.50 Å	A/B	448-723	PDBe RCSB ... PDBj PDBsu...
2FB8	X-ray	2.90 Å	A/B	445-723	PDBe

https://www.uniprot.org/uniprot/P15056/protvista

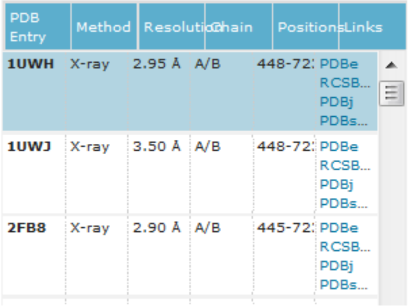
UniProtKB annotation & Feature viewer



UniProtKB - P15056 (BRAF_HUMAN)

NP_BIND 463-471

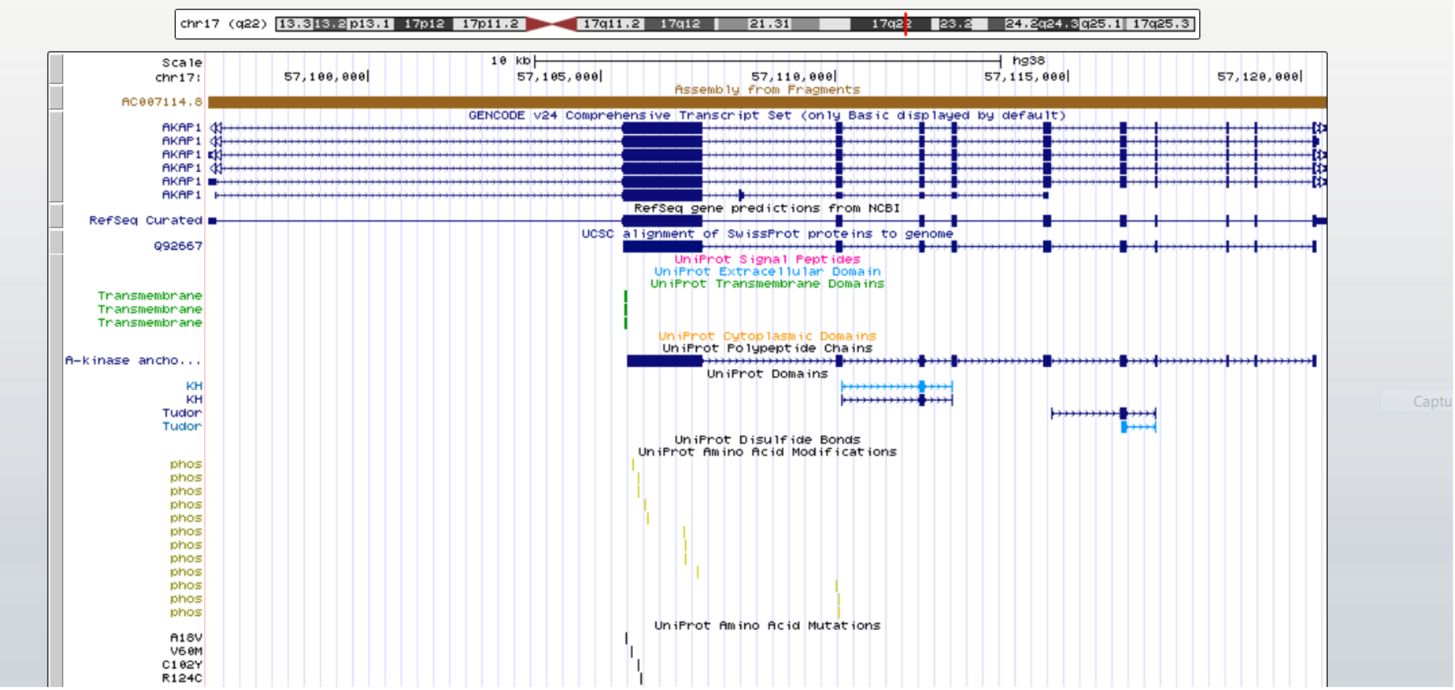
Description	ATP
Evidence	PROSITE-ProRule annotation (Manual assertion according to rules)
	PROSITE-ProRule PRU00159
Tools	BLAST

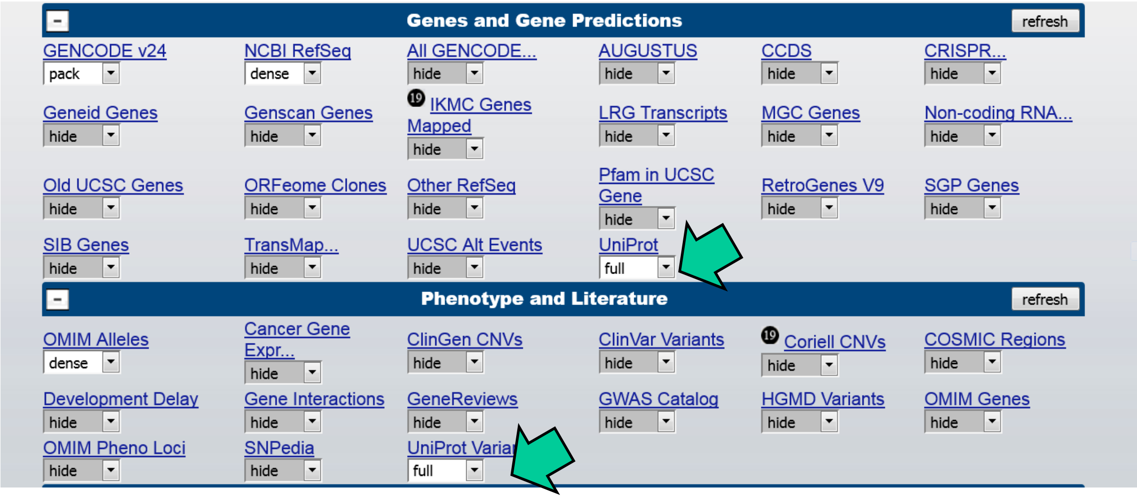


PDB Entry	Method	Resolutio	Chain	Positions	Links
1UWH	X-ray	2.95 Å	A/B	448-72:	PDBe RCSB... PDBj PDBs...
1UWJ	X-ray	3.50 Å	A/B	448-72:	PDBe RCSB... PDBj PDBs...
2FB8	X-ray	2.90 Å	A/B	445-72:	PDBe RCSB... PDBj PDBs...

UniProtKB annotation & Genome browsers

http://www.uniprot.org/uniprot/Q92667


chr17 (q22)
13.3 13.2 p13.1 17p12 17p11.2 17q11.2 17q12 21.31 17q22 23.2 24.2 q24.3 q25.1 17q25.3
Scale
chr17:
10 kb
hg38
57,100,000
57,105,000
57,110,000
57,115,000
57,120,000
Assembly from Fragments
AC007114.8
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
AKAP1
AKAP1
AKAP1
AKAP1
AKAP1
AKAP1
RefSeq gene predictions from NCBI
RefSeq Curated
UCSC alignment of SwissProt proteins to genome
Q92667
UniProt Signal Peptides
UniProt Extracellular Domain
UniProt Transmembrane Domains
Transmembrane
Transmembrane
Transmembrane
UniProt Cytoplasmic Domains
UniProt Polypeptide Chains
A-kinase ancho...
UniProt Domains
KH
KH
Tudor
Tudor
UniProt Disulfide Bonds
UniProt Amino Acid Modifications
phos
phos
phos
phos
phos
phos
phos
phos
phos
phos
phos
phos
UniProt Amino Acid Mutations
A18V
V60M
C102Y
R124C
Captu



Genes and Gene Predictions

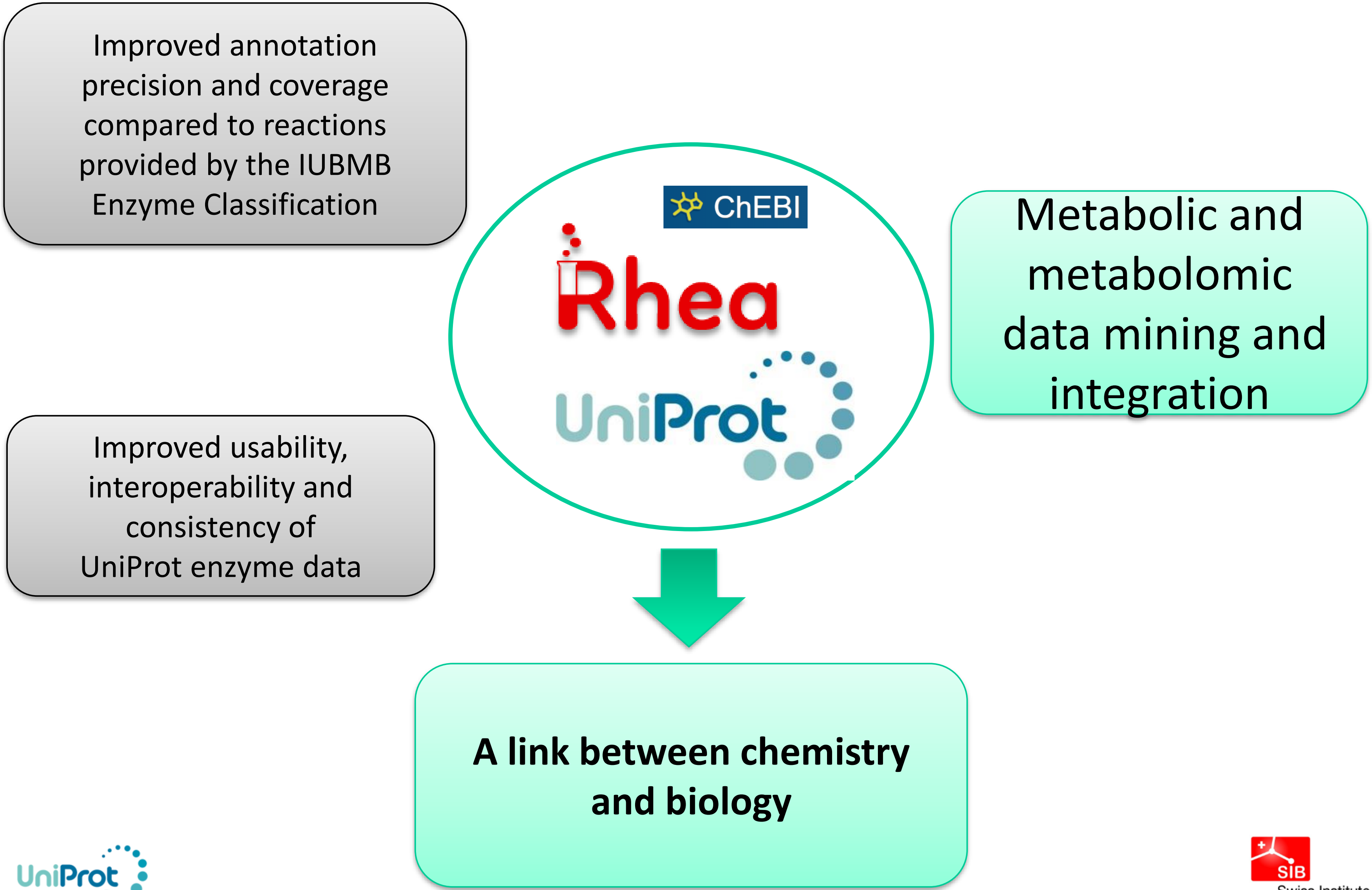
GENCODE v24: pack
NCBI RefSeq: dense
All GENCODE...: hide
AUGUSTUS: hide
CCDS: hide
CRISPR...: hide
Geneid Genes: hide
Genscan Genes: hide
IKMC Genes Mapped: hide
LRG Transcripts: hide
MGC Genes: hide
Non-coding RNA...: hide
Old UCSC Genes: hide
ORFeome Clones: hide
Other RefSeq: hide
Pfam in UCSC Gene: hide
RetroGenes V9: hide
SGP Genes: hide
SIB Genes: hide
TransMap...: hide
UCSC Alt Events: hide
UniProt: full

Phenotype and Literature

OMIM Alleles: dense
Cancer Gene Expr...: hide
ClinGen CNVs: hide
ClinVar Variants: hide
Coriell CNVs: hide
COSMIC Regions: hide
Development Delay: hide
Gene Interactions: hide
GeneReviews: hide
GWAS Catalog: hide
HGMD Variants: hide
OMIM Genes: hide
OMIM Pheno Loci: hide
SNPedia: hide
UniProt Varia: full

The UniProt release 2018_11 will be publicly available on **December 5th**.

Rhea in UniProt



Improved annotation precision and coverage compared to reactions provided by the IUBMB Enzyme Classification

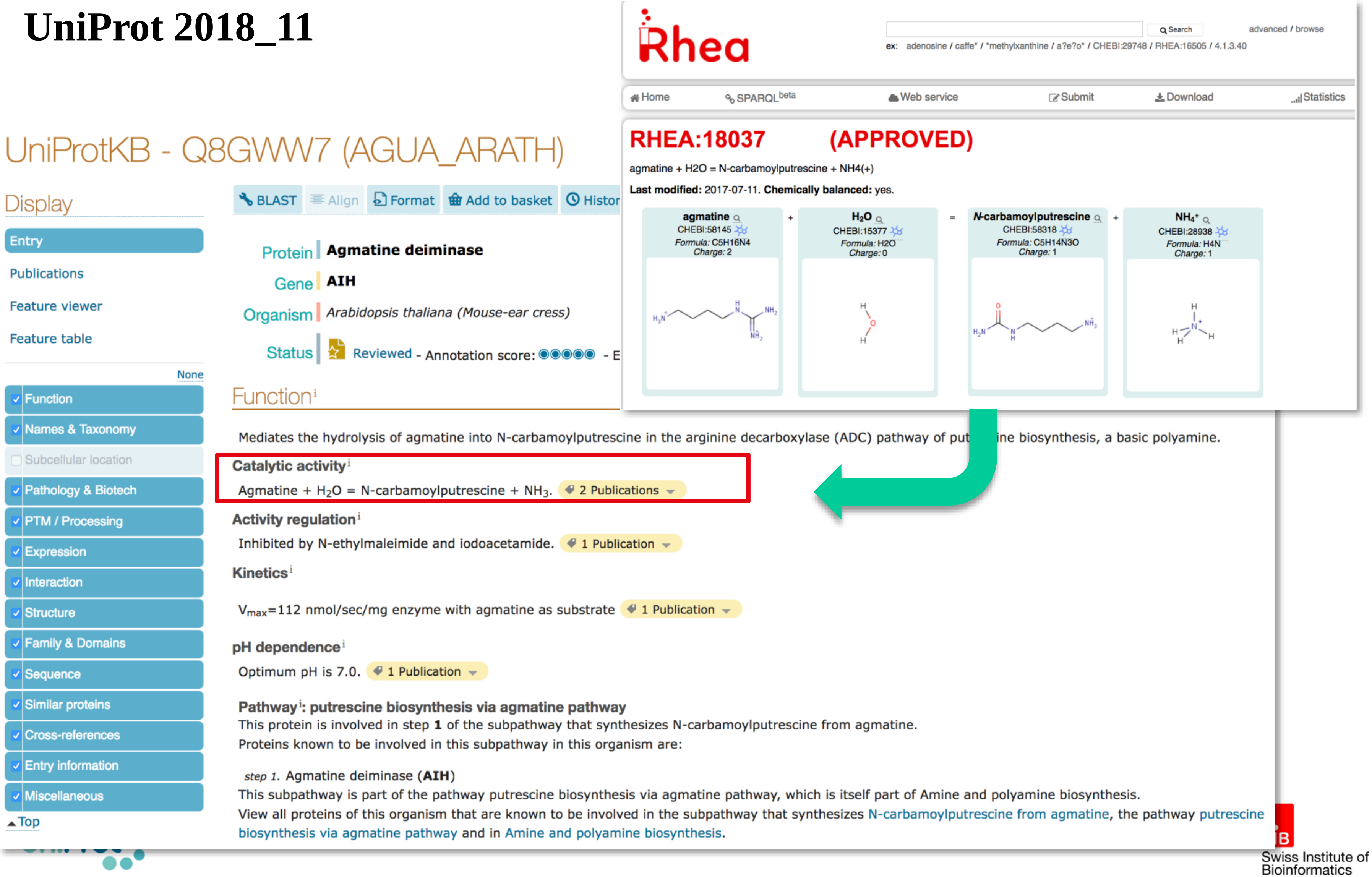
Improved usability, interoperability and consistency of UniProt enzyme data

Metabolic and metabolomic data mining and integration

A link between chemistry and biology

What will change?

UniProt 2018_11



Rhea

Search advanced / browse

ex: adenosine / caffe* / *methylxanthine / a?e?o* / CHEBI:29748 / RHEA:16505 / 4.1.3.40

Home | SPARQL beta | Web service | Submit | Download | Statistics

RHEA:18037 (APPROVED)

agmatine + H2O = N-carbamoylputrescine + NH4(+)

Last modified: 2017-07-11. **Chemically balanced:** yes.

agmatine	+	H_2O	=	*N*-carbamoylputrescine	+	NH_4^+
CHEBI:58145		CHEBI:15377		CHEBI:58318		CHEBI:28938
Formula: C5H16N4		*Formula:* H2O		*Formula:* C5H14N3O		*Formula:* H4N
Charge: 2		*Charge:* 0		*Charge:* 1		*Charge:* 1

UniProtKB - Q8GWW7 (AGUA_ARATH)

Display

- Entry
- Publications
- Feature viewer
- Feature table

None

- ☑ Function
- ☑ Names & Taxonomy
- □ Subcellular location
- ☑ Pathology & Biotech
- ☑ PTM / Processing
- ☑ Expression
- ☑ Interaction
- ☑ Structure
- ☑ Family & Domains
- ☑ Sequence
- ☑ Similar proteins
- ☑ Cross-references
- ☑ Entry information
- ☑ Miscellaneous

▲Top

BLAST | Align | Format | Add to basket | Histor

Protein	**Agmatine deiminase**
Gene	**AIH**
Organism	*Arabidopsis thaliana (Mouse-ear cress)*
Status	Reviewed - Annotation score: ●●●●● - E

Function

Mediates the hydrolysis of agmatine into N-carbamoylputrescine in the arginine decarboxylase (ADC) pathway of put ine biosynthesis, a basic polyamine.

Catalytic activity

Agmatine + H_2O = N-carbamoylputrescine + NH_3. 2 Publications

Activity regulation

Inhibited by N-ethylmaleimide and iodoacetamide. 1 Publication

Kinetics

V_{max}=112 nmol/sec/mg enzyme with agmatine as substrate 1 Publication

pH dependence

Optimum pH is 7.0. 1 Publication

Pathway: putrescine biosynthesis via agmatine pathway

This protein is involved in step **1** of the subpathway that synthesizes N-carbamoylputrescine from agmatine.

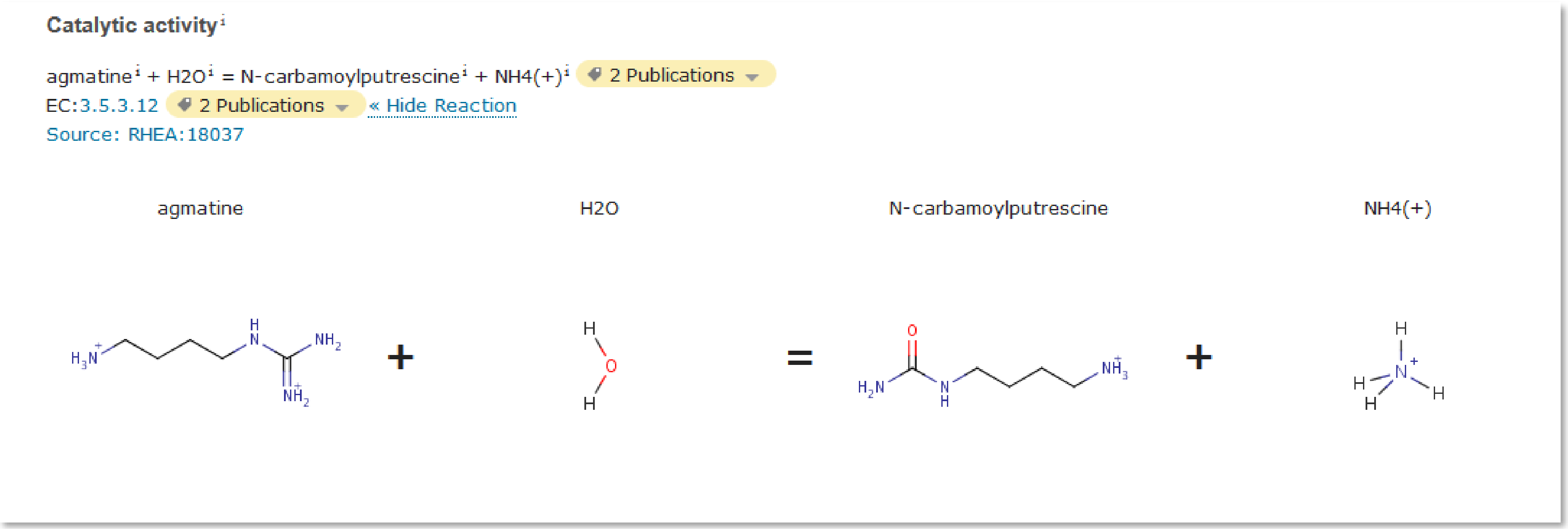
Proteins known to be involved in this subpathway in this organism are:

step 1. Agmatine deiminase (**AIH**)

This subpathway is part of the pathway putrescine biosynthesis via agmatine pathway, which is itself part of Amine and polyamine biosynthesis.

View all proteins of this organism that are known to be involved in the subpathway that synthesizes N-carbamoylputrescine from agmatine, the pathway putrescine biosynthesis via agmatine pathway and in Amine and polyamine biosynthesis.

Display on the UniProt website



Catalytic activity

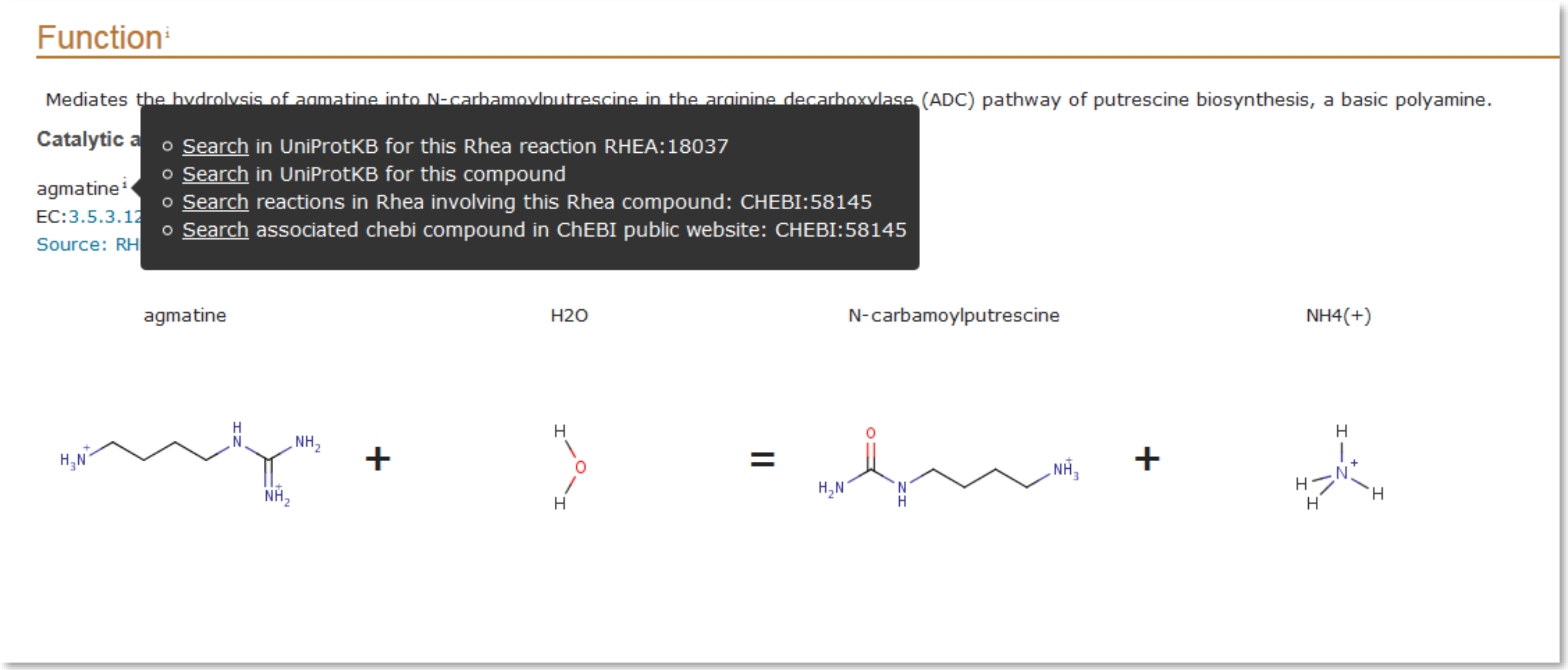
agmatine + H2O = N-carbamoylputrescine + NH4(+) 2 Publications

EC:3.5.3.12 2 Publications « Hide Reaction

Source: RHEA:18037

Display on the UniProt website

- **Under development…**



Function

Mediates the hydrolysis of agmatine into N-carbamoylputrescine in the arginine decarboxylase (ADC) pathway of putrescine biosynthesis, a basic polyamine.

Catalytic a

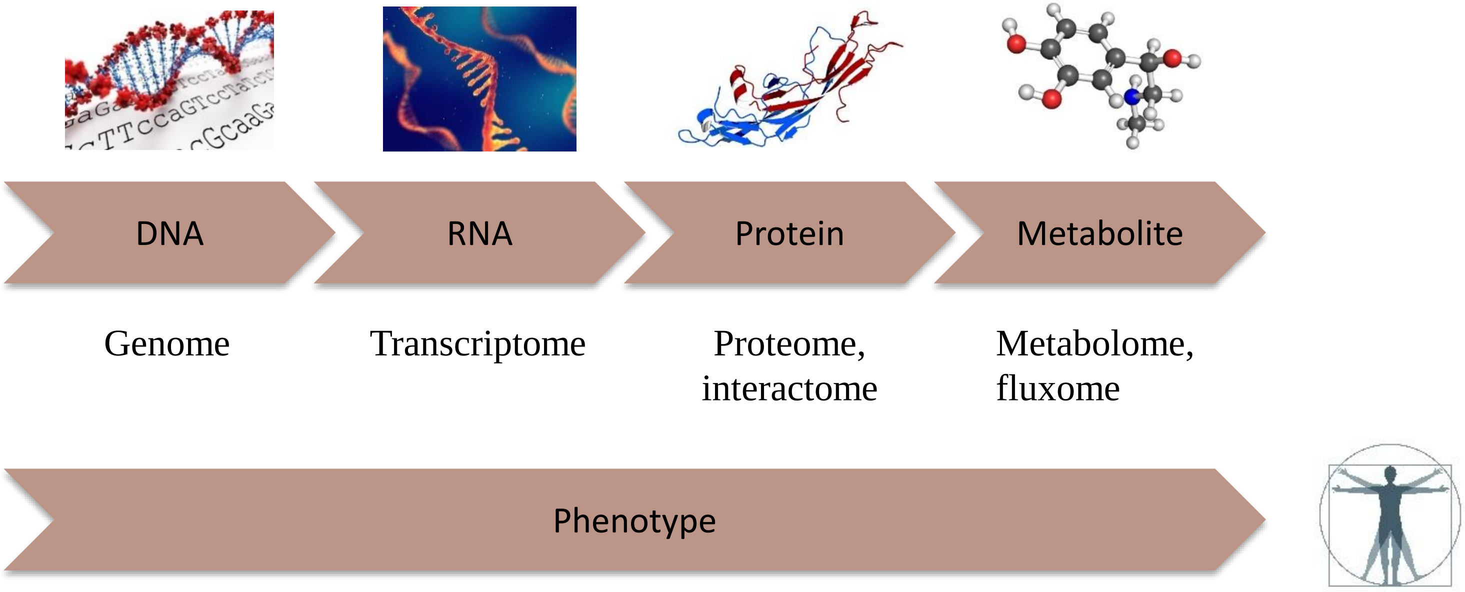
agmatine

EC:3.5.3.12

Source: RH

- ○ Search in UniProtKB for this Rhea reaction RHEA:18037
- ○ Search in UniProtKB for this compound
- ○ Search reactions in Rhea involving this Rhea compound: CHEBI:58145
- ○ Search associated chebi compound in ChEBI public website: CHEBI:58145

A complete integration of data



DNA

RNA

Protein

Metabolite

Genome

Transcriptome

Proteome,
interactome

Metabolome,
fluxome

Phenotype

A key to understand metabolic phenotype

What kind of searches are now possible ?

Searches on the UniProt website or SPARQL queries

- What are the human proteins that catalyze reactions using D-glucose ? *(a specific compound)*
- What are the enzymes in species X involved in lipid metabolism ? *(a class of compounds)*
- What is a reaction network for a specified organism of interest ?
- Retrieve the links between genes, transcripts and proteins to relevant metabolites, in a specified organism
- Identify putative enzymes acting on a specific metabolite

- Nucleic acid sequence databases
    - INSDC, RefSeq, Ensembl

- Protein sequence databases

- UniProtKB
    - UniProtKB/Swiss-Prot
        - Protein sequences
        - Biological knowledge
    - UniProtKB/TrEMBL
        - Protein sequences
        - Biological knowledge
    - → GO annotation
    - UniProt Proteomes

- NCBI protein (RefSeq)

- Practicals

Gene Ontology

- The Gene Ontology is a controlled vocabulary, a set of standard terms—words and phrases—used for indexing and retrieving information.

- Created by the GO consortium

- Contains ~50’000 terms.

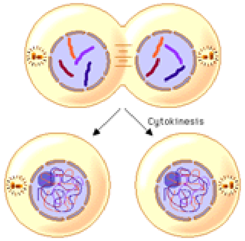
- GO also defines the relationships between the terms (hierarchy), making it a structured vocabulary.

- **`http://www.geneontology.org`**

3 categories of GO terms

1. Biological Process

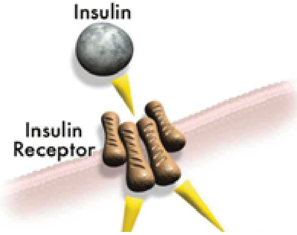
A commonly recognized series of events



- Cell division
- Mitosis
- Organelle fission

2. Molecular Function

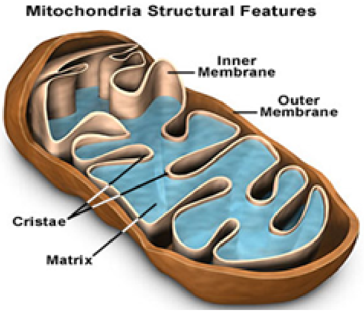
An elemental activity or task or job



- Protein kinase activity
- Insulin binding
- Insulin receptor activity

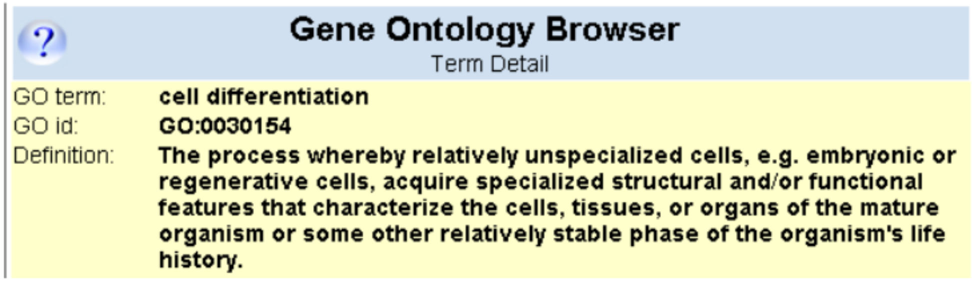
3. Cellular Component

Where a gene product is located



- Mitochondrion
- Mitochondrial matrix
- Mitochondrial membrane

Each GO term has two definitions

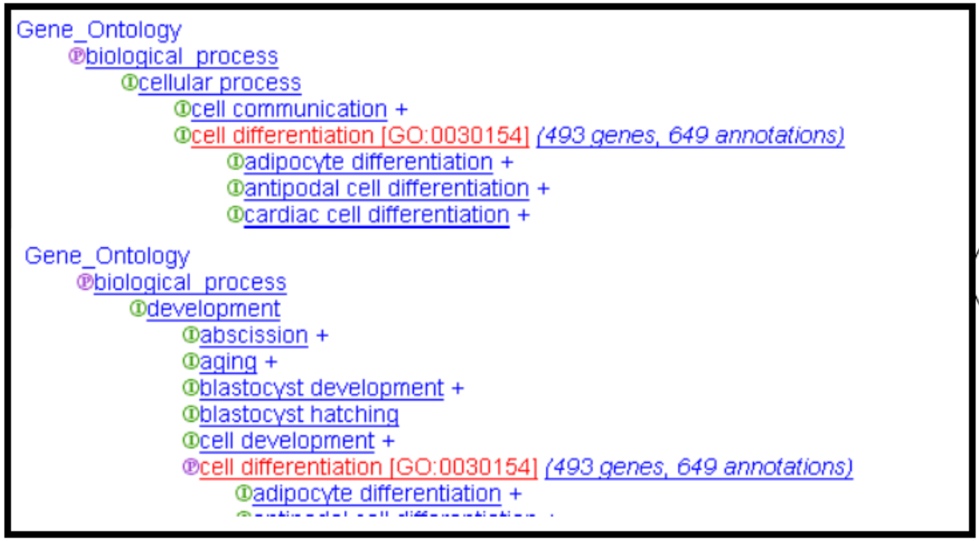


Gene Ontology Browser

Term Detail

GO term:	**cell differentiation**
GO id:	**GO:0030154**
Definition:	**The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.**

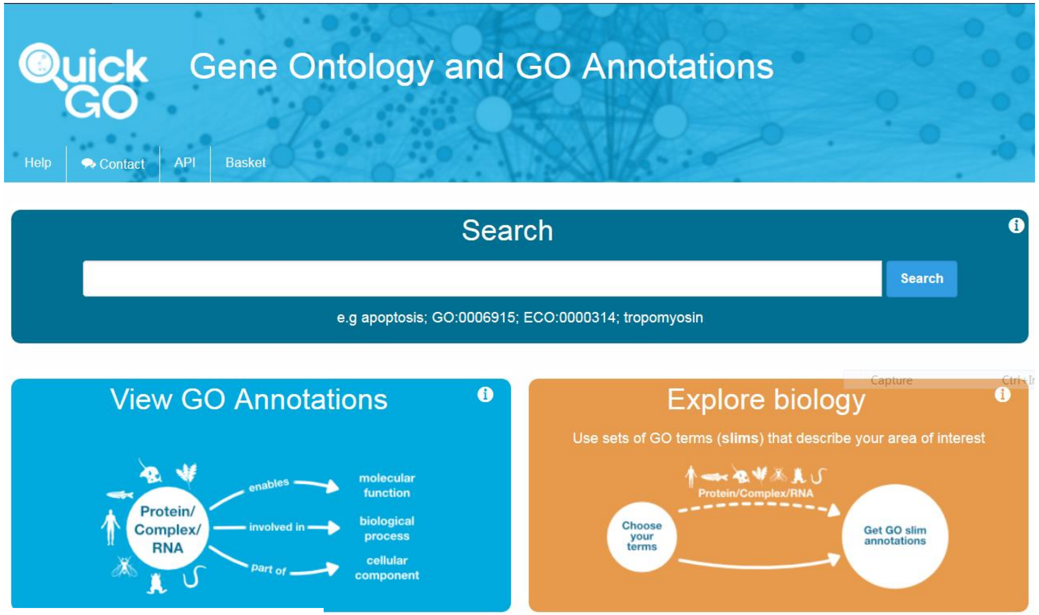
A textual definition
written by
a biologist



```
Gene_Ontology
    Ⓟbiological_process
        Ⓘcellular process
            Ⓘcell communication +
            Ⓘcell differentiation [GO:0030154] (493 genes, 649 annotations)
                Ⓘadipocyte differentiation +
                Ⓘantipodal cell differentiation +
                Ⓘcardiac cell differentiation +

Gene_Ontology
    Ⓟbiological_process
        Ⓘdevelopment
            Ⓘabscission +
            Ⓘaging +
            Ⓘblastocyst development +
            Ⓘblastocyst hatching
            Ⓘcell development +
            Ⓟcell differentiation [GO:0030154] (493 genes, 649 annotations)
                Ⓘadipocyte differentiation +
                Ⓘantipodal cell differentiation +
```

Graph structure
(~hierarchy): **formal
and computable**

GO databases



Gene Ontology and GO Annotations

Search

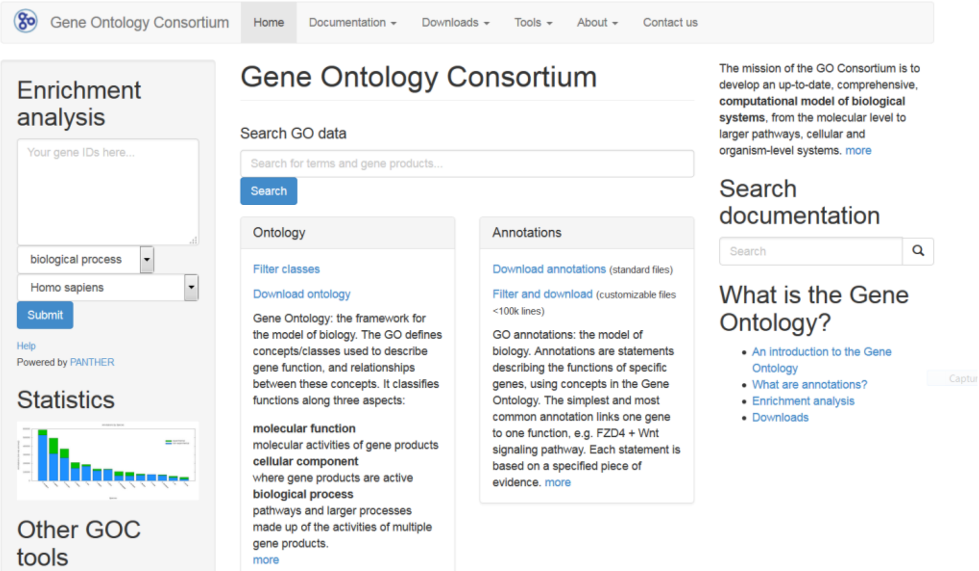
Search

e.g apoptosis; GO:0006915; ECO:0000314; tropomyosin

Explore biology

Use sets of GO terms (**slims**) that describe your area of interest

http://www.ebi.ac.uk/QuickGO/



Enrichment analysis

Help

Powered by PANTHER

Statistics

Other GOC tools

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Ontology

Filter classes

Download ontology

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

molecular function
molecular activities of gene products
cellular component
where gene products are active
biological process
pathways and larger processes made up of the activities of multiple gene products.
more

Annotations

Download annotations (standard files)

Filter and download (customizable files <100k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. more

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. more

Search documentation

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Enrichment analysis
- Downloads

http://www.geneontology.org

SIB Swiss Institute of Bioinformatics

Gene Ontology (GO) annotation

UniProt

SIB Swiss Institute of Bioinformatics

GO Annotation

1. **Experimental annotations (EXP)**

   Biocurators read papers and summarize the content with a list of appropriated GO terms

2. **Curated non experimental annotation**

   Evidence such as sequence similarity, database entries

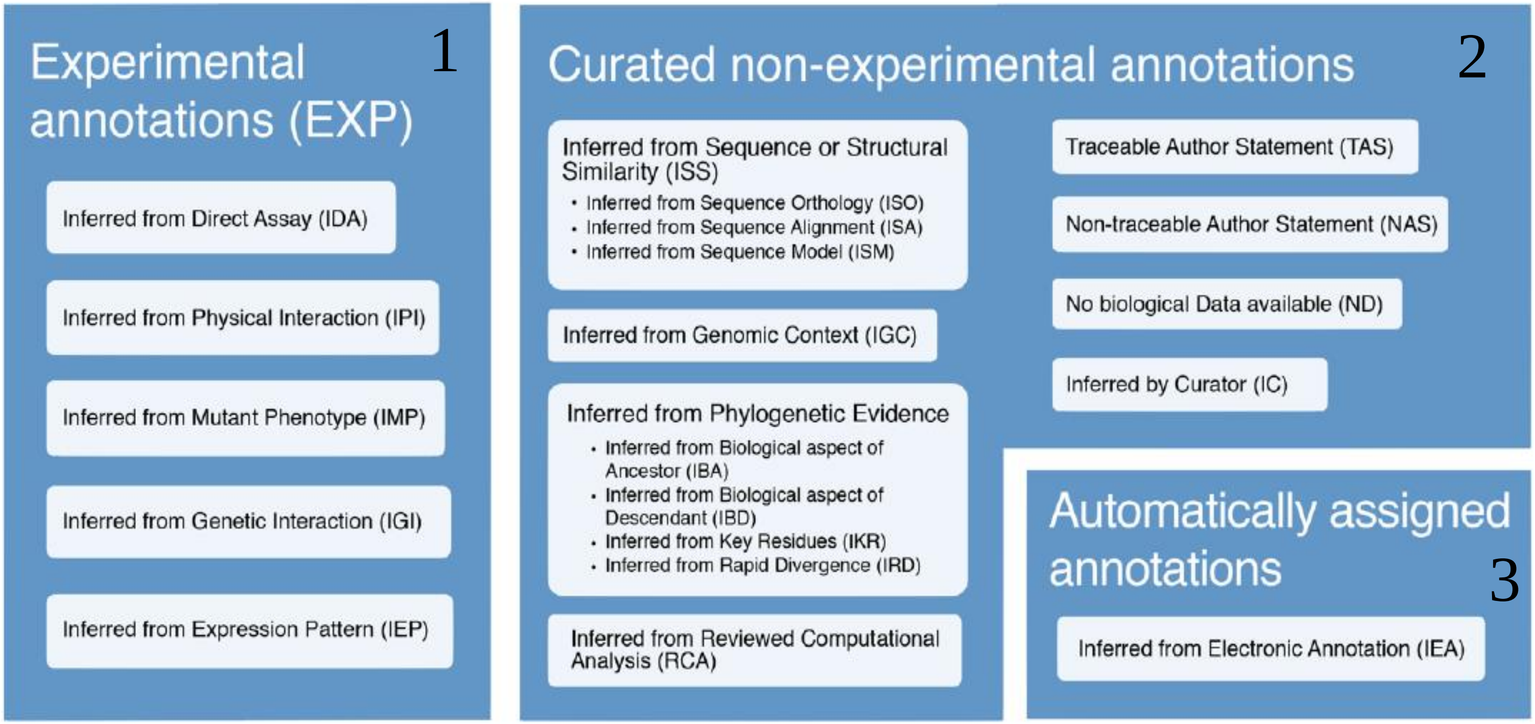
3. **Automatically assigned annotation (IEA)**

   No curator oversight

   Inferred by Electronic Annotation: **> 98 %**

Swiss Institute of Bioinformatics

Annotations are supported by Evidence codes



Experimental annotations (EXP) 1

- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)

Curated non-experimental annotations 2

- Inferred from Sequence or Structural Similarity (ISS)
  - Inferred from Sequence Orthology (ISO)
  - Inferred from Sequence Alignment (ISA)
  - Inferred from Sequence Model (ISM)
- Inferred from Genomic Context (IGC)
- Inferred from Phylogenetic Evidence
  - Inferred from Biological aspect of Ancestor (IBA)
  - Inferred from Biological aspect of Descendant (IBD)
  - Inferred from Key Residues (IKR)
  - Inferred from Rapid Divergence (IRD)
- Inferred from Reviewed Computational Analysis (RCA)
- Traceable Author Statement (TAS)
- Non-traceable Author Statement (NAS)
- No biological Data available (ND)
- Inferred by Curator (IC)

Automatically assigned annotations 3

- Inferred from Electronic Annotation (IEA)

>98 %

Figure from: Gaudet, Skunka, Hu, Dessimoz, The Gene Ontology Handbook 2017

GO annotation: experimental (manual)

PUBMED 1: Protein kinase R (PKR) recognizes double stranded RNA in the cytoplasm

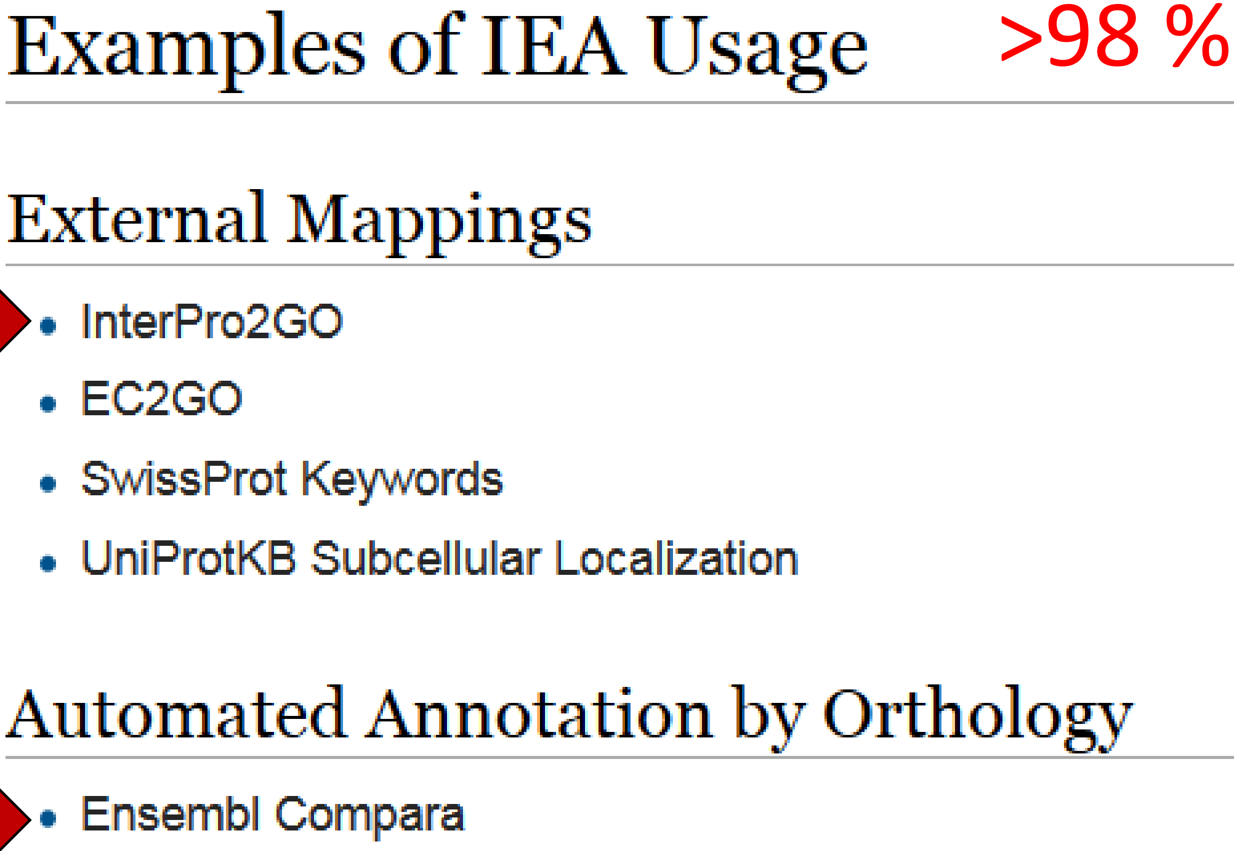
PUBMED 2: PKR acts as a kinase in the nucleus

<u>GO annotation</u>:

Molecular Function: -double stranded RNA binding
-kinase activity

Cellular Component: -nucleus
-cytoplasm

GO annotation: automated



>98 %

Examples of IEA Usage

External Mappings

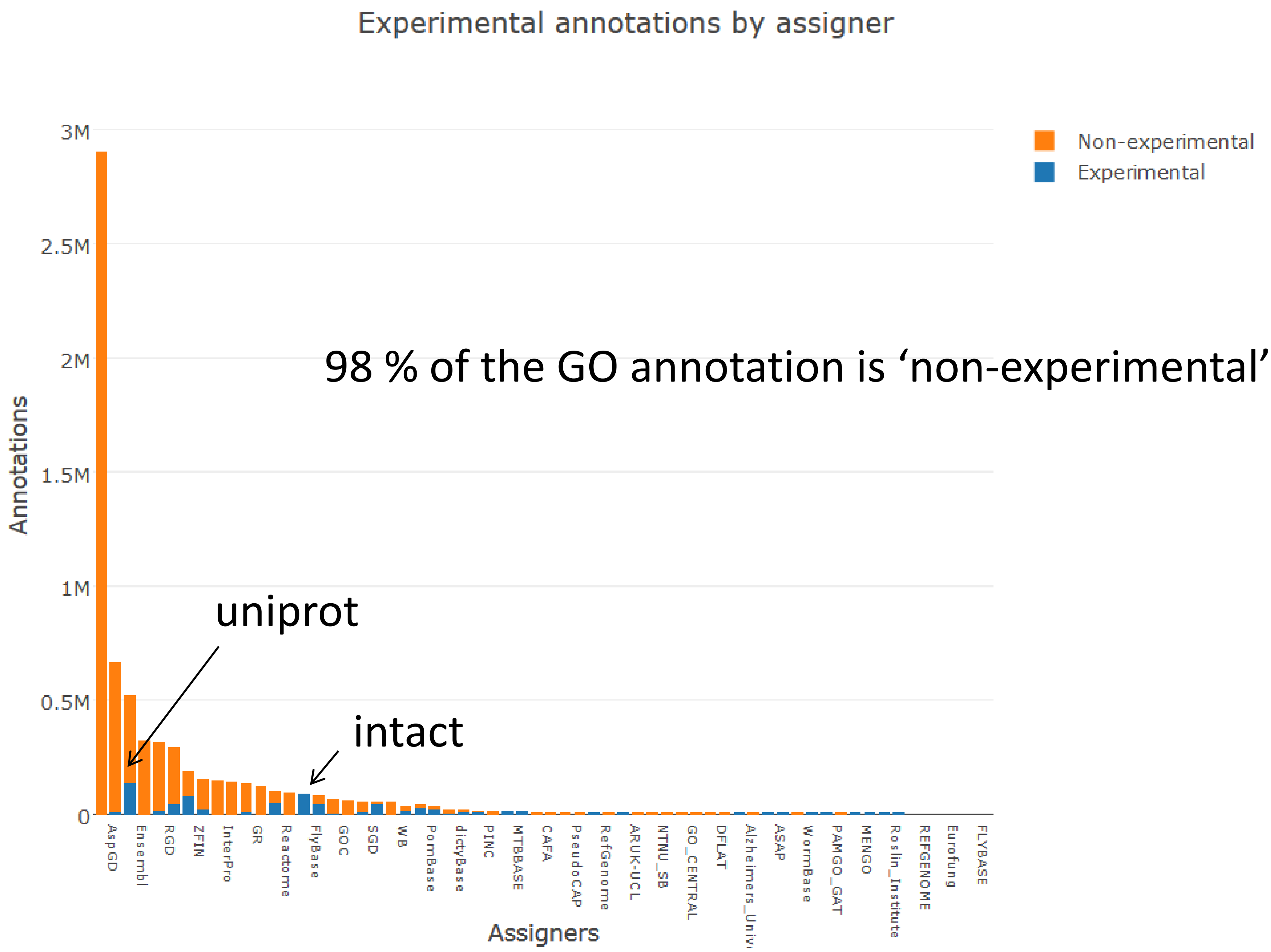
- InterPro2GO
- EC2GO
- SwissProt Keywords
- UniProtKB Subcellular Localization

Automated Annotation by Orthology

- Ensembl Compara

UniProt

SIB Swiss Institute of Bioinformatics



98 % of the GO annotation is ‘non-experimental’

uniprot

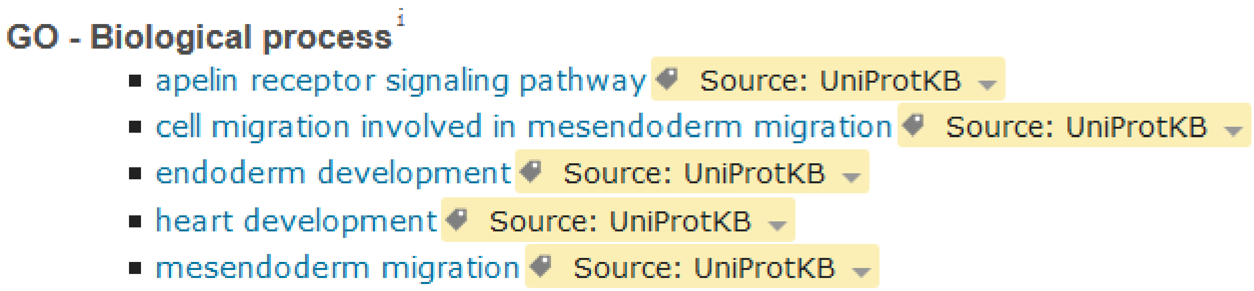
intact

UniProt

SIB Swiss Institute of Bioinformatics

(1) You can have different set of GO terms for a same gene depending on the database

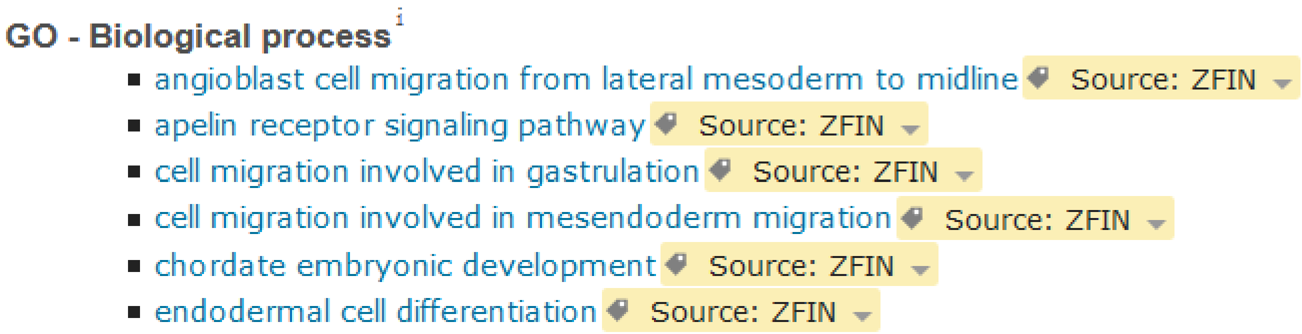
UniProtKB/Swiss-Prot: P0DMC2 (ELA_DANRE)



GO - Biological process

- apelin receptor signaling pathway Source: UniProtKB
- cell migration involved in mesendoderm migration Source: UniProtKB
- endoderm development Source: UniProtKB
- heart development Source: UniProtKB
- mesendoderm migration Source: UniProtKB

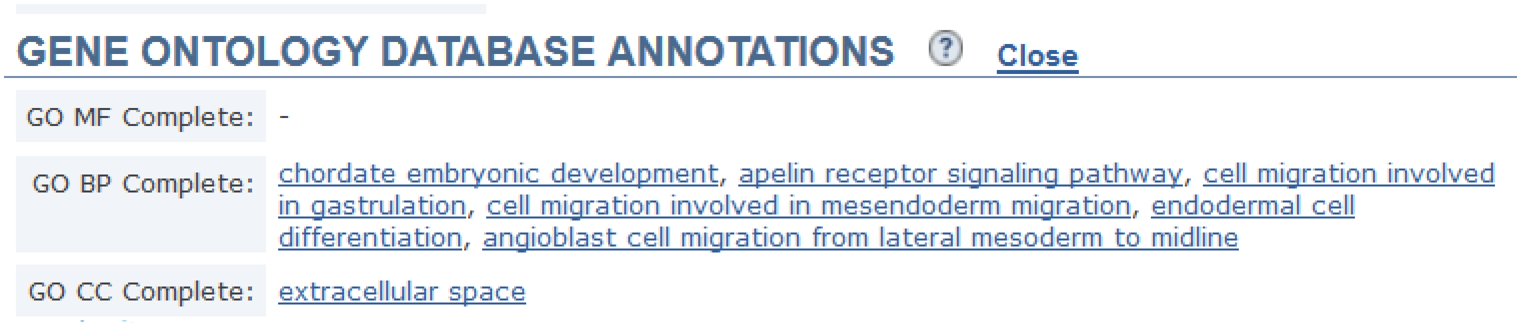
UniProtKB/TrEMBL: A0A0R4IR01_DANRE



GO - Biological process

- angioblast cell migration from lateral mesoderm to midline Source: ZFIN
- apelin receptor signaling pathway Source: ZFIN
- cell migration involved in gastrulation Source: ZFIN
- cell migration involved in mesendoderm migration Source: ZFIN
- chordate embryonic development Source: ZFIN
- endodermal cell differentiation Source: ZFIN

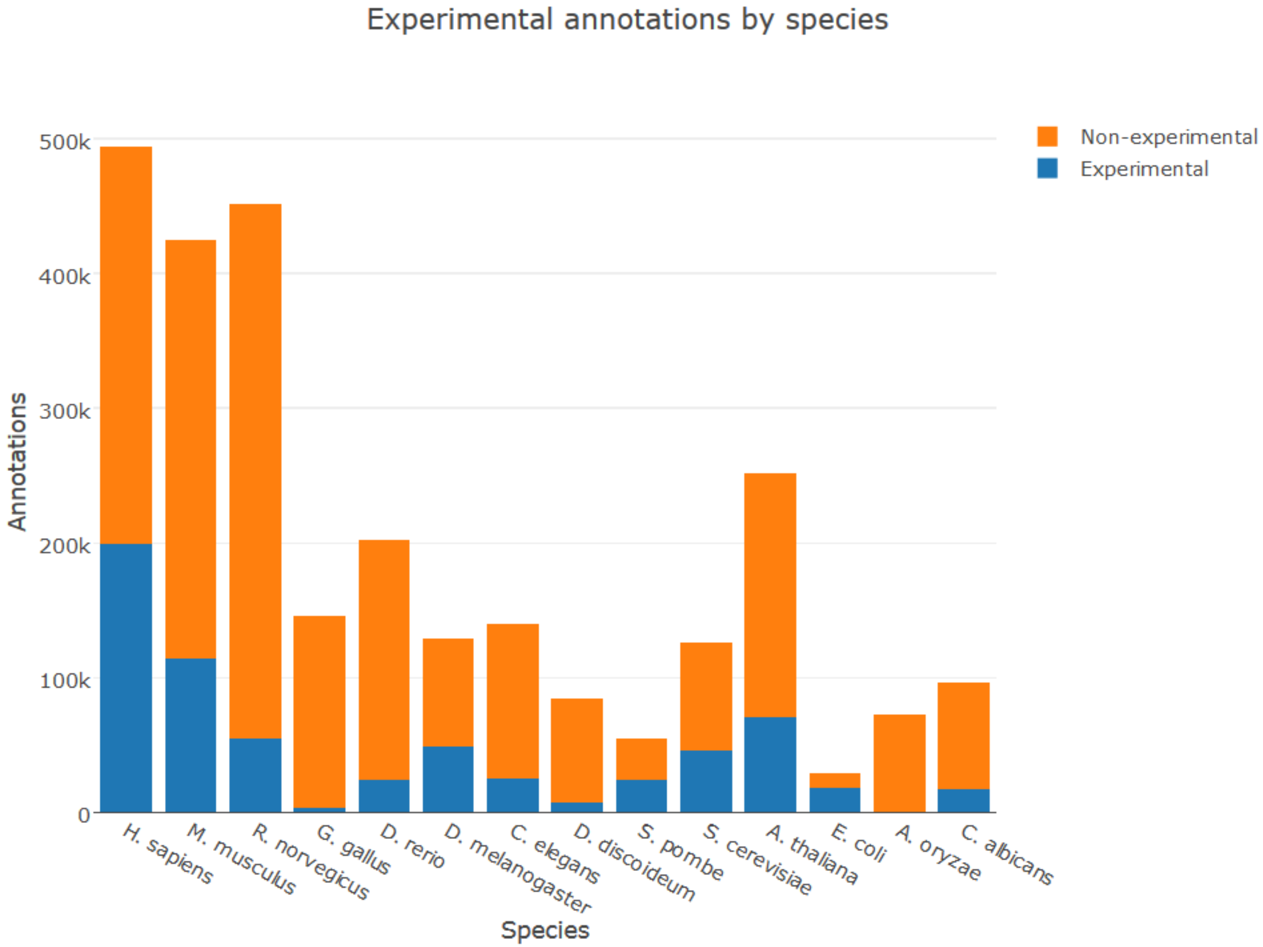
Panther: P0DMC2 (ELA_DANRE)



GENE ONTOLOGY DATABASE ANNOTATIONS

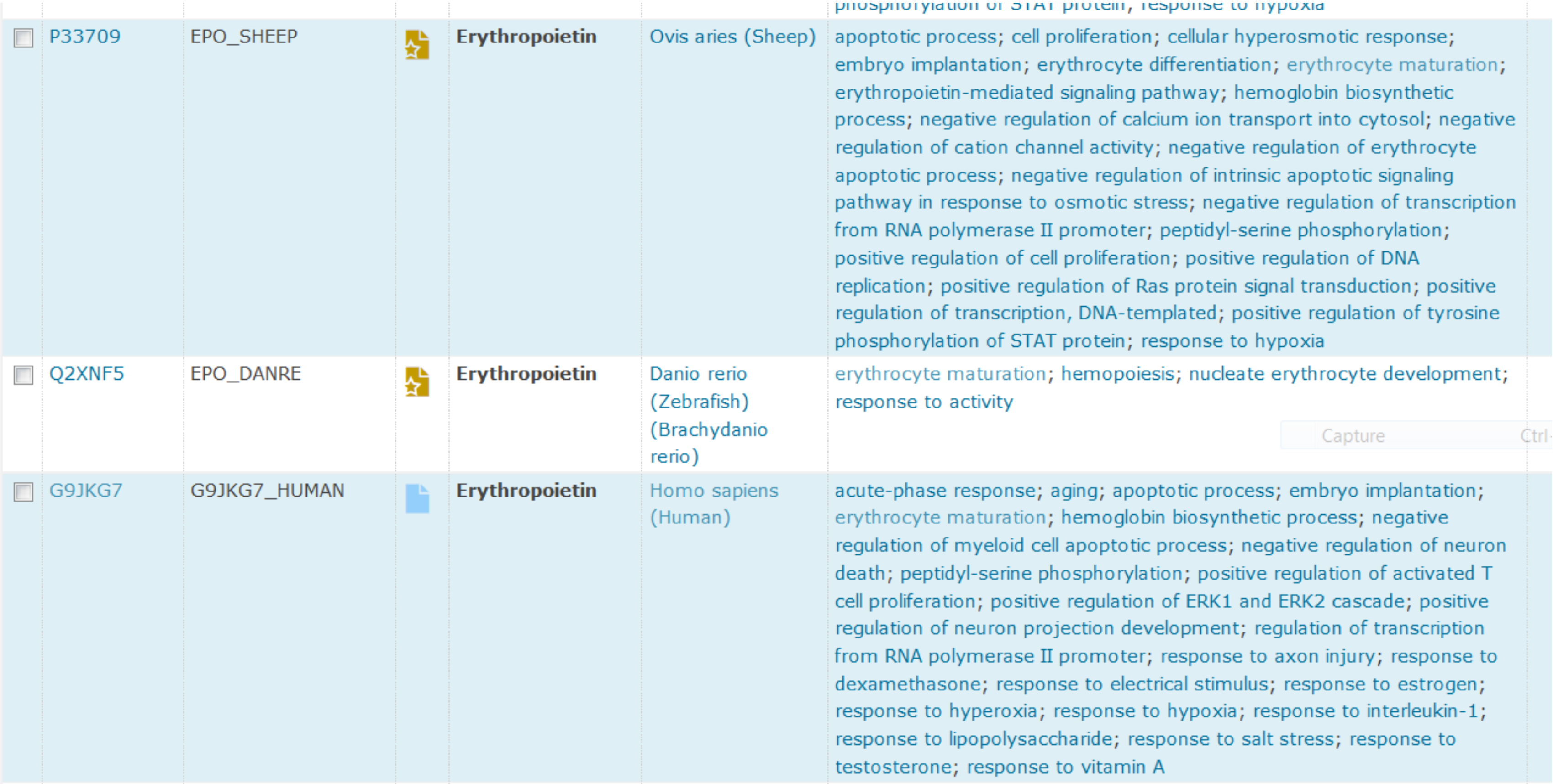
GO MF Complete:	-
GO BP Complete:	chordate embryonic development, apelin receptor signaling pathway, cell migration involved in gastrulation, cell migration involved in mesendoderm migration, endodermal cell differentiation, angioblast cell migration from lateral mesoderm to midline
GO CC Complete:	extracellular space

(2) The ‘quantity’ and ‘quality’ of GO annotation varie from one species to another



http://amigo.geneontology.org/amigo/base_statistics

(2) The ‘quantity’ and ‘quality’ of GO annotation varie from one species to another



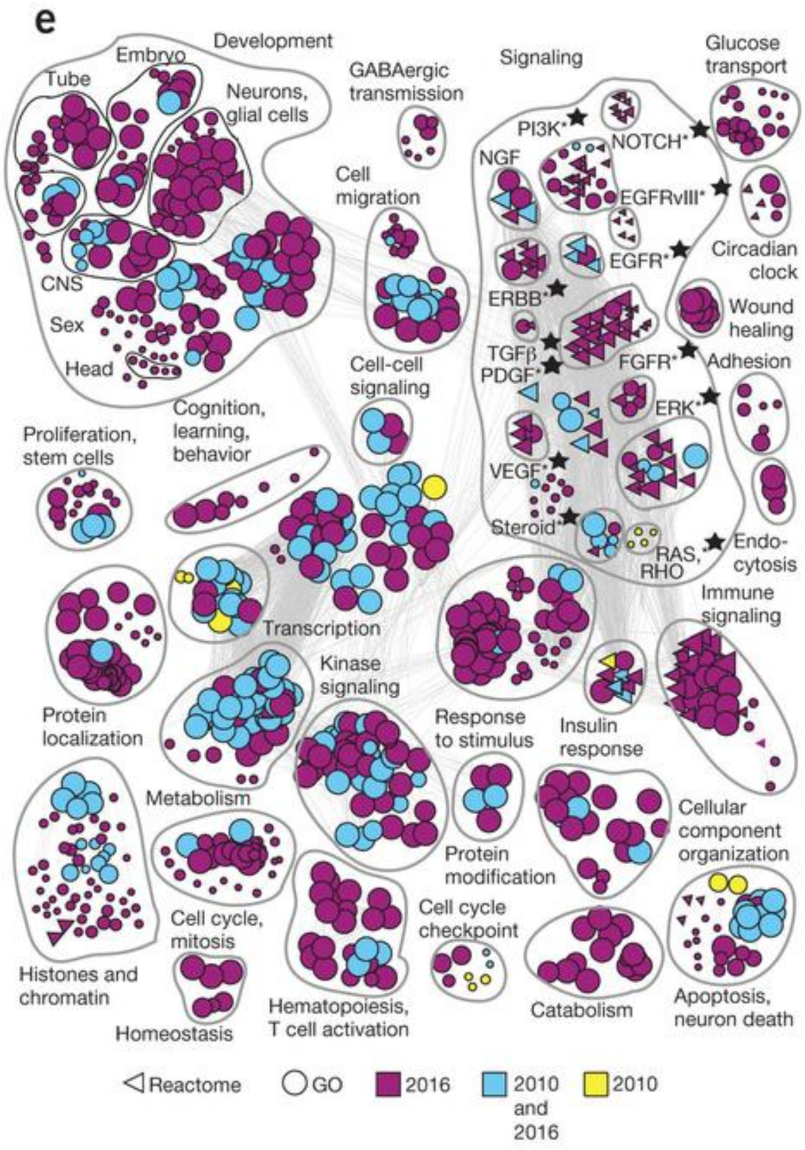
						phosphorylation of STAT protein; response to hypoxia
□	P33709	EPO_SHEEP	★	**Erythropoietin**	Ovis aries (Sheep)	apoptotic process; cell proliferation; cellular hyperosmotic response; embryo implantation; erythrocyte differentiation; erythrocyte maturation; erythropoietin-mediated signaling pathway; hemoglobin biosynthetic process; negative regulation of calcium ion transport into cytosol; negative regulation of cation channel activity; negative regulation of erythrocyte apoptotic process; negative regulation of intrinsic apoptotic signaling pathway in response to osmotic stress; negative regulation of transcription from RNA polymerase II promoter; peptidyl-serine phosphorylation; positive regulation of cell proliferation; positive regulation of DNA replication; positive regulation of Ras protein signal transduction; positive regulation of transcription, DNA-templated; positive regulation of tyrosine phosphorylation of STAT protein; response to hypoxia
□	Q2XNF5	EPO_DANRE	★	**Erythropoietin**	Danio rerio (Zebrafish) (Brachydanio rerio)	erythrocyte maturation; hemopoiesis; nucleate erythrocyte development; response to activity
□	G9JKG7	G9JKG7_HUMAN	🗎	**Erythropoietin**	Homo sapiens (Human)	acute-phase response; aging; apoptotic process; embryo implantation; erythrocyte maturation; hemoglobin biosynthetic process; negative regulation of myeloid cell apoptotic process; negative regulation of neuron death; peptidyl-serine phosphorylation; positive regulation of activated T cell proliferation; positive regulation of ERK1 and ERK2 cascade; positive regulation of neuron projection development; regulation of transcription from RNA polymerase II promoter; response to axon injury; response to dexamethasone; response to electrical stimulus; response to estrogen; response to hyperoxia; response to hypoxia; response to interleukin-1; response to lipopolysaccharide; response to salt stress; response to testosterone; response to vitamin A

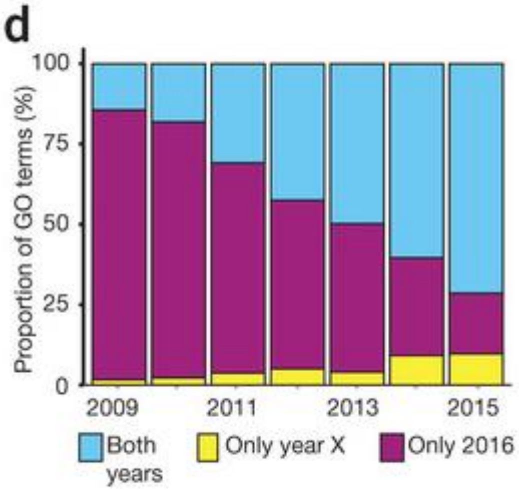
Where do they come from ?

UniProt

SIB
Swiss Institute of Bioinformatics

(3) timeliness of GO annotation…





Impact of outdated gene annotations on pathway enrichment analysis

(August 2016)

Analysis of 75 mutated glioblastoma (GBM) genes using annual annotations from 2009–2016

Pathway analysis assesses the statistical enrichment of biological processes and pathways in a given gene list on the basis of information in Gene Ontology (GO) and pathway databases such as Reactome and PathwayCommons. GO is updated daily and Reactome versions are released quarterly, but many software tools interpret gene lists using functional information that has not been updated for years.

http://www.nature.com/nmeth/journal/v13/n9/full/nmeth.3963.html

Nucleic acid sequence databases

- INSDC, RefSeq, Ensembl

Protein sequence databases

UniProtKB

- UniProtKB/Swiss-Prot
  - Protein sequences
  - Biological knowledge
- UniProtKB/TrEMBL
  - Protein sequences
  - Biological knowledge
- GO annotation
- ➡ UniProt Proteomes

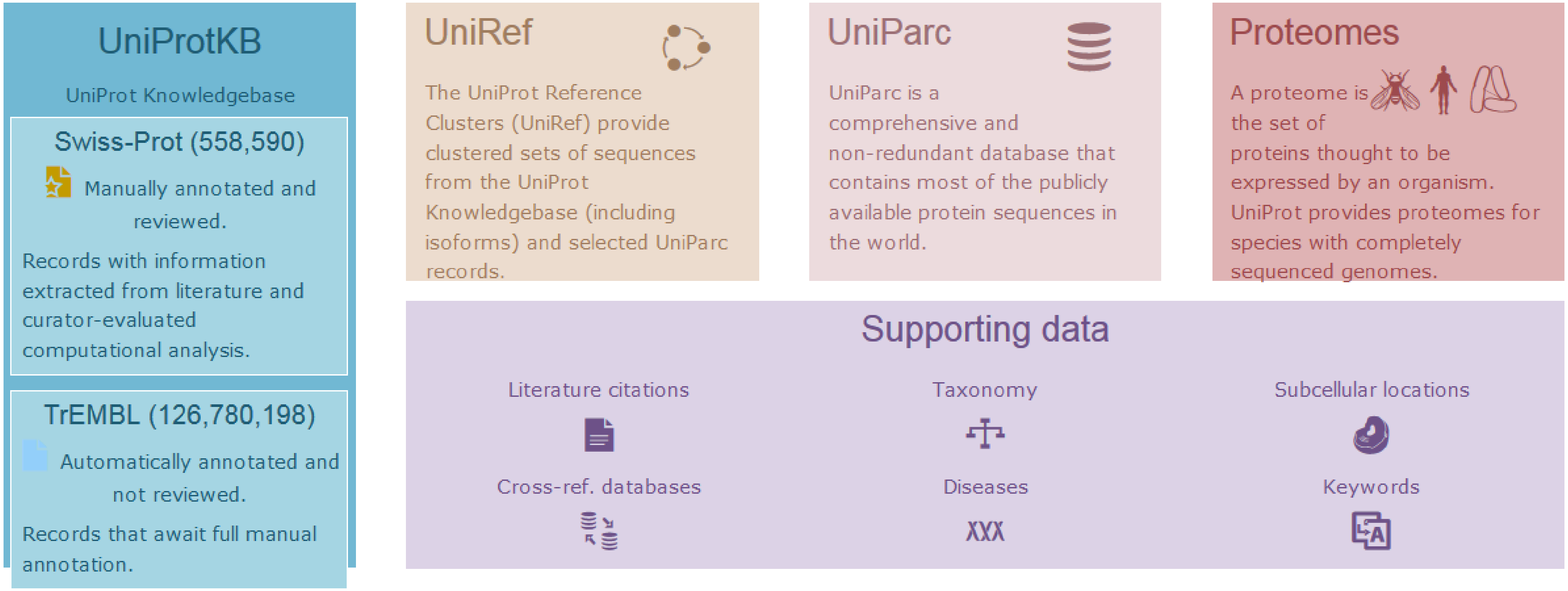
NCBI protein (RefSeq)

Practicals

UniProt

SIB Swiss Institute of Bioinformatics

UniProtKB Proteomes



UniProtKB

UniProt Knowledgebase

Swiss-Prot (558,590)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (126,780,198)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations

Taxonomy

Subcellular locations

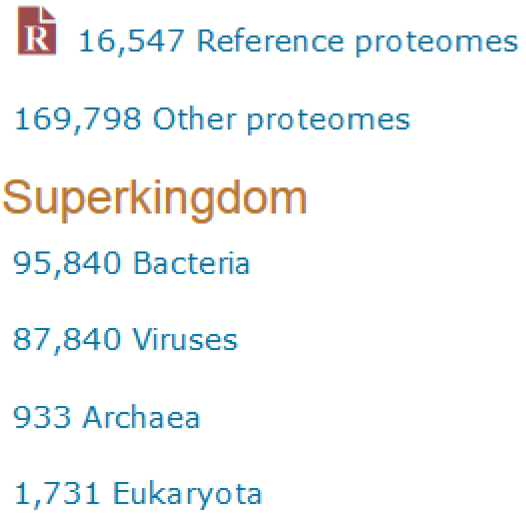
Cross-ref. databases

Diseases

Keywords

Proteomes / Reference proteomes

- A **proteome** is the set of proteins thought to be expressed by an organism (completely sequenced genomes).
- **A proteome is formed from all UniProtKB/Swiss-Prot entries + those UniProtKB/TrEMBL entries mapping to Ensembl (genomes).**
- Pseudogenes and other dubious uncharacterized ORFs are removed



R 16,547 Reference proteomes

169,798 Other proteomes

Superkingdom

95,840 Bacteria

87,840 Viruses

933 Archaea

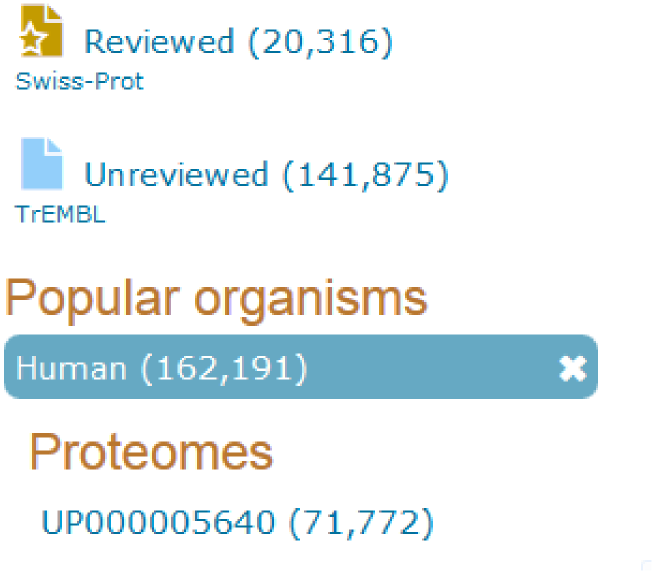
1,731 Eukaryota

UniProt

SIB Swiss Institute of Bioinformatics

Human proteins

(records)



Reviewed (20,316)
Swiss-Prot

Unreviewed (141,875)
TrEMBL

Popular organisms

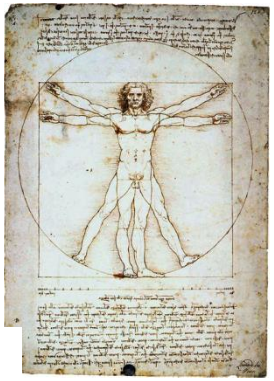
Human (162,191) ✖

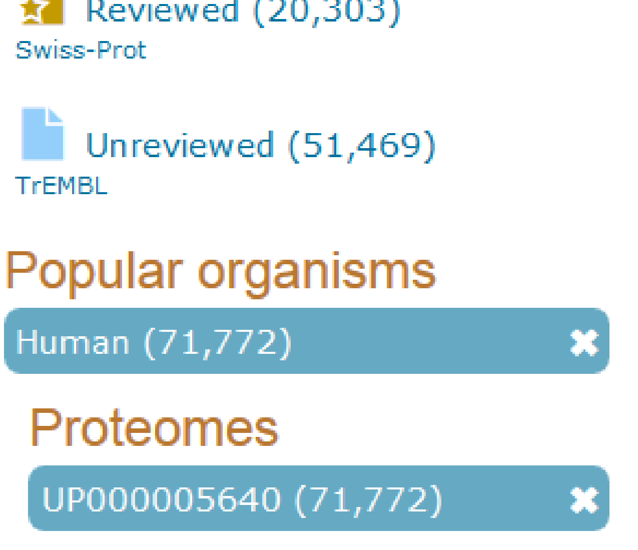
Proteomes

UP000005640 (71,772)

Query: organism:"Homo sapiens (Human) [9606]"

Human proteome





Reviewed (20,303)
Swiss-Prot

Unreviewed (51,469)
TrEMBL

Popular organisms

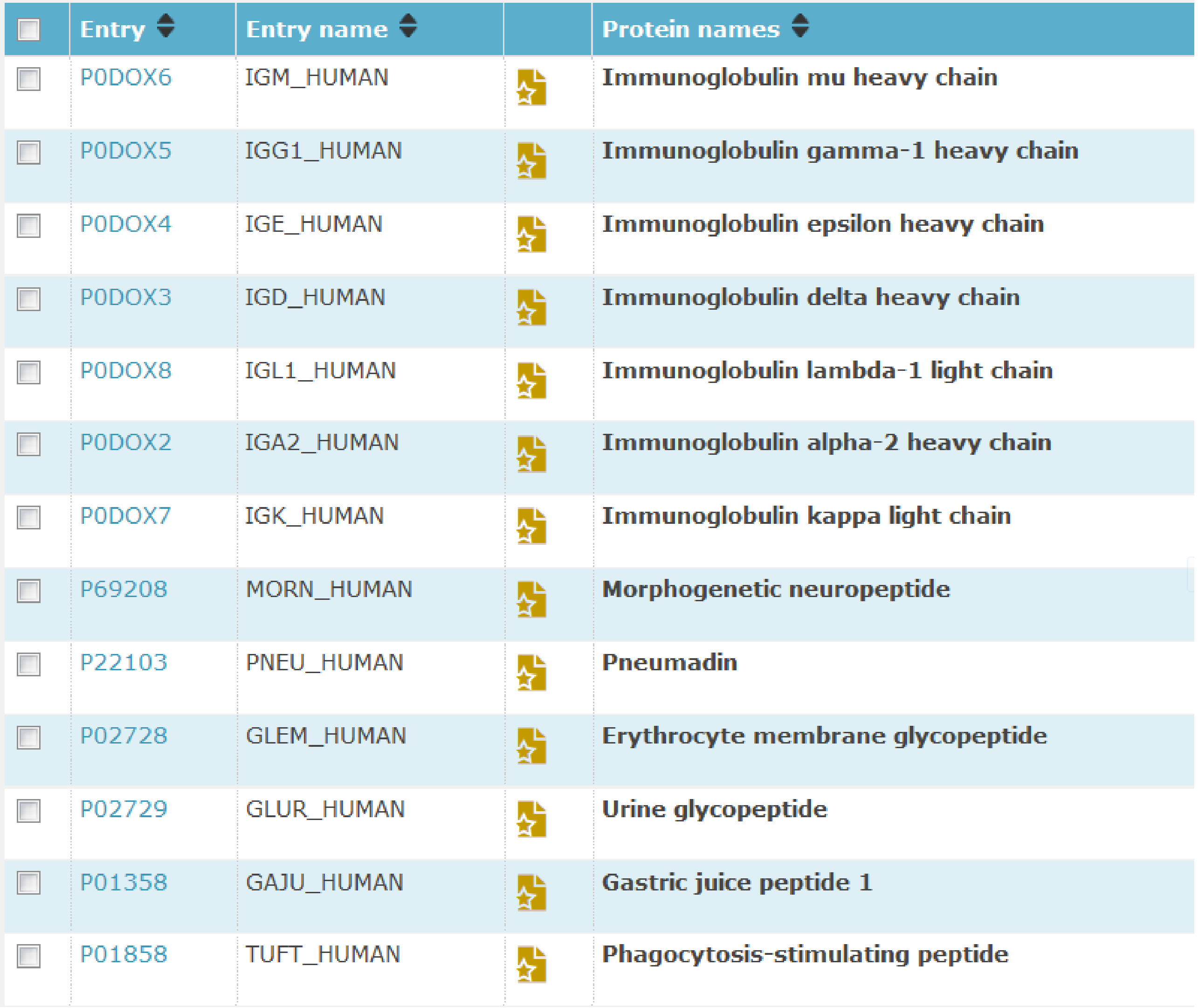
Human (71,772) ✖

Proteomes

UP000005640 (71,772) ✖

Query: organism:"Homo sapiens (Human) [9606]" AND proteome:up000005640

+ Swiss-Prot isoforms: ~22 000 (link)



□	Entry	Entry name		Protein names
□	P0DOX6	IGM_HUMAN	★	**Immunoglobulin mu heavy chain**
□	P0DOX5	IGG1_HUMAN	★	**Immunoglobulin gamma-1 heavy chain**
□	P0DOX4	IGE_HUMAN	★	**Immunoglobulin epsilon heavy chain**
□	P0DOX3	IGD_HUMAN	★	**Immunoglobulin delta heavy chain**
□	P0DOX8	IGL1_HUMAN	★	**Immunoglobulin lambda-1 light chain**
□	P0DOX2	IGA2_HUMAN	★	**Immunoglobulin alpha-2 heavy chain**
□	P0DOX7	IGK_HUMAN	★	**Immunoglobulin kappa light chain**
□	P69208	MORN_HUMAN	★	**Morphogenetic neuropeptide**
□	P22103	PNEU_HUMAN	★	**Pneumadin**
□	P02728	GLEM_HUMAN	★	**Erythrocyte membrane glycopeptide**
□	P02729	GLUR_HUMAN	★	**Urine glycopeptide**
□	P01358	GAJU_HUMAN	★	**Gastric juice peptide 1**
□	P01858	TUFT_HUMAN	★	**Phagocytosis-stimulating peptide**

NOT proteome:up000005640 AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

Swiss Institute of Bioinformatics

Proteomes / Reference proteomes

- Some proteomes have been (manually and algorithmically) selected as **reference proteomes** (useful for biomedical research and phylogeny).

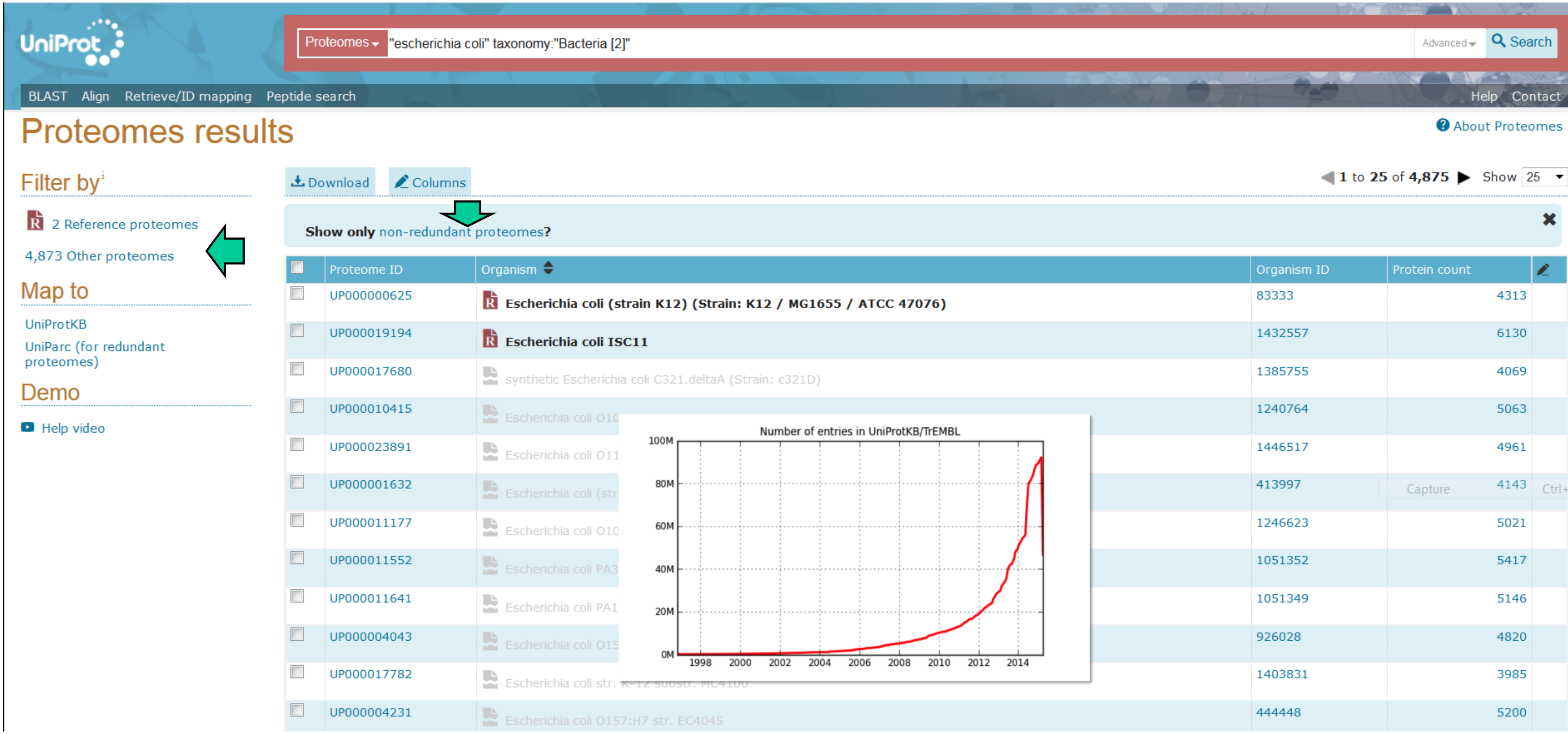
- Regularly updated

⚠ http://www.uniprot.org/help/proteome
http://www.uniprot.org/help/reference_proteome

UniProt

SIB Swiss Institute of Bioinformatics

Proteomes / Reference proteomes



Proteomes results

Filter by

- 2 Reference proteomes
- 4,873 Other proteomes

Map to

- UniProtKB
- UniParc (for redundant proteomes)

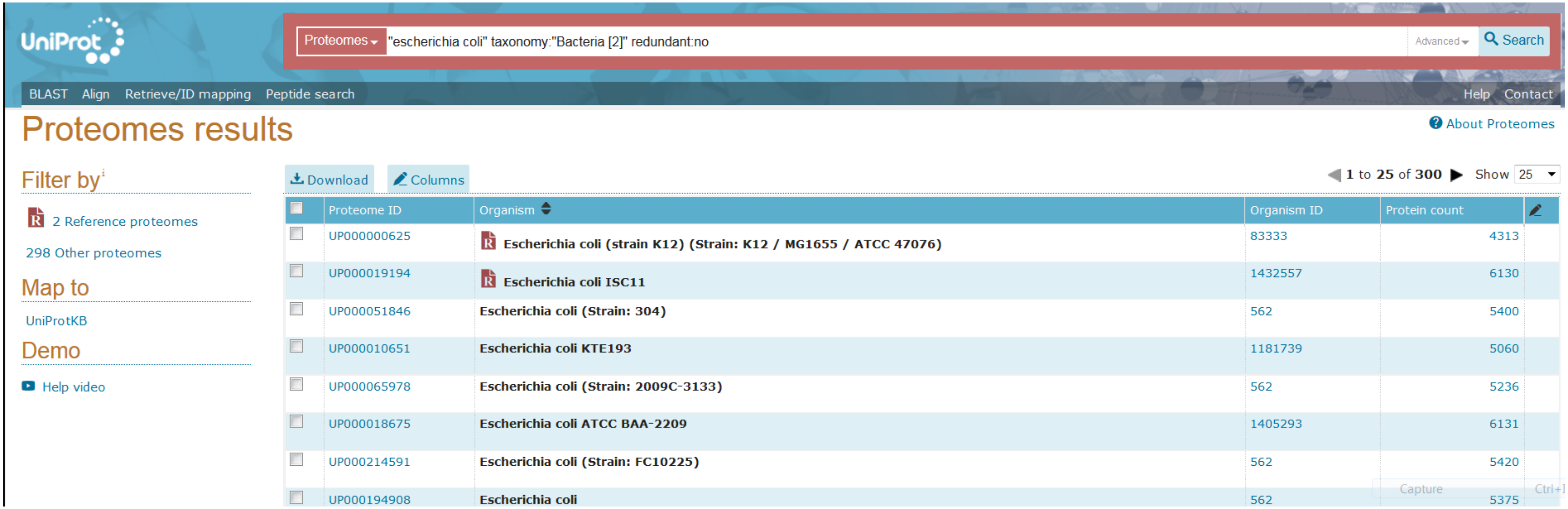
Demo

- Help video

Show only non-redundant proteomes?

☐	Proteome ID	Organism	Organism ID	Protein count
☐	UP000000625	**Escherichia coli (strain K12) (Strain: K12 / MG1655 / ATCC 47076)**	83333	4313
☐	UP000019194	**Escherichia coli ISC11**	1432557	6130
☐	UP000017680	synthetic Escherichia coli C321.deltaA (Strain: c321D)	1385755	4069
☐	UP000010415	Escherichia coli O10	1240764	5063
☐	UP000023891	Escherichia coli O11	1446517	4961
☐	UP000001632	Escherichia coli (str	413997	4143
☐	UP000011177	Escherichia coli O10	1246623	5021
☐	UP000011552	Escherichia coli PA3	1051352	5417
☐	UP000011641	Escherichia coli PA1	1051349	5146
☐	UP000004043	Escherichia coli O15	926028	4820
☐	UP000017782	Escherichia coli str. K-12 substr. MC4100	1403831	3985
☐	UP000004231	Escherichia coli O157:H7 str. EC4045	444448	5200

The ‘redundant’ proteomes are excluded from UniProtKB (since 2015).
The corresponding protein sequences are only available in UniParc …



"escherichia coli" taxonomy:"Bacteria [2]" redundant:no

Proteomes results

About Proteomes

Filter by

R 2 Reference proteomes

298 Other proteomes

Map to

UniProtKB

Demo

Help video

1 to 25 of 300

□	Proteome ID	Organism	Organism ID	Protein count
□	UP000000625	R Escherichia coli (strain K12) (Strain: K12 / MG1655 / ATCC 47076)	83333	4313
□	UP000019194	R Escherichia coli ISC11	1432557	6130
□	UP000051846	Escherichia coli (Strain: 304)	562	5400
□	UP000010651	Escherichia coli KTE193	1181739	5060
□	UP000065978	Escherichia coli (Strain: 2009C-3133)	562	5236
□	UP000018675	Escherichia coli ATCC BAA-2209	1405293	6131
□	UP000214591	Escherichia coli (Strain: FC10225)	562	5420
□	UP000194908	Escherichia coli	562	5375

Proteomes redundancy is only applied to bacteria and fungi for the moment.
See FAQ:
https://www.uniprot.org/help/redundancy
https://www.uniprot.org/help/proteome_redundancy
http://insideuniprot.blogspot.ch/2015_05_01_archive.html

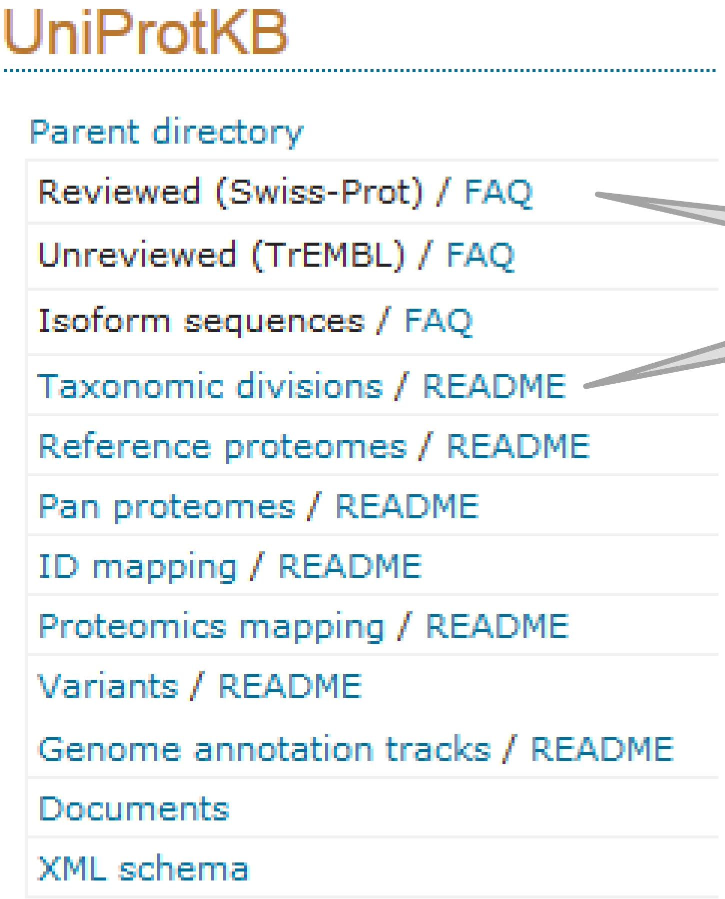
If you need ‘to protect a proteome for a good reason, please contact Andrea.Auchincloss@sib.swiss or Ivo.Pedruzzi@sib.swiss

UniProt

SIB Swiss Institute of Bioinformatics

Downloads

Downloads



UniProtKB

Parent directory

- Reviewed (Swiss-Prot) / FAQ
- Unreviewed (TrEMBL) / FAQ
- Isoform sequences / FAQ
- Taxonomic divisions / README
- Reference proteomes / README
- Pan proteomes / README
- ID mapping / README
- Proteomics mapping / README
- Variants / README
- Genome annotation tracks / README
- Documents
- XML schema

Flat file (.dat)

Flat file

Flat file

Flat file

Fasta format (.fasta)

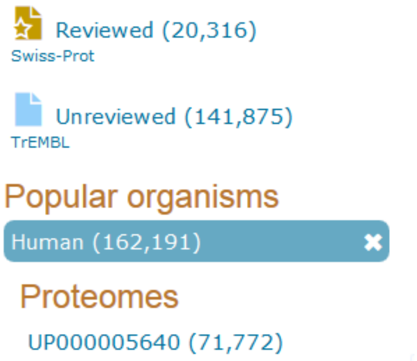
Fasta format

Does not include «Swiss-Prot isoforms»

http://www.uniprot.org/downloads

Query + Download: homo sapiens

updated



Reviewed (20,316)
Swiss-Prot

Unreviewed (141,875)
TrEMBL

Popular organisms

Human (162,191)

Proteomes

UP000005640 (71,772)

HUMAN (TaxID: 9606) WEB	Swiss-Prot (canonical)	Swiss-Prot (isoforms)	TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (without Swiss-Prot isoforms)
Total	20 316*	22 014	141 875#	184 205	162 191
Proteome	20 303	22 014	51 469	**93 786**	71 772

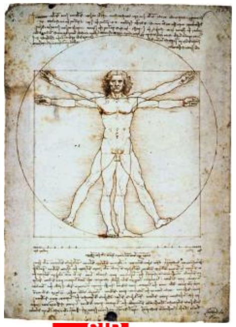
*uniprot_sprot_human.dat.gz #uniprot_trembl_human.dat.gz

HUMAN FTP (README) Reference proteome	Main fasta Swiss-Prot & TrEMBL (« canonical »)	Additional fasta Swiss-Prot & TrEMBL (« isoforms »)	Swiss-Prot & TrEMBL (canonical + isoforms)
Total	20 998*	72 788#	**93 786**

*UP000005640_9606.fasta.gz

#UP000005640_9606_additional.fasta.gz

UniProt release 2018_02

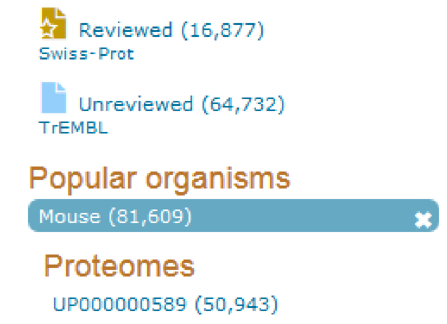


UniProt

* # ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/README

SIB Swiss Institute of Bioinformatics

Query + Download: mus musculus


Reviewed (16,877)
Swiss-Prot

Unreviewed (64,732)
TrEMBL

Popular organisms

Mouse (81,609)

Proteomes

UP000000589 (50,943)

MOUSE (TaxID: 10090) WEB	Swiss-Prot (canonical)	Swiss-Prot (isoforms)	TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (without Swiss-Prot isoforms)
Total	16 877	8 140	64 732	89 749	81 609
Proteome	16 873	8 138	34 070	**59 081**	50 943

MOUSE FTP (README) Reference proteome	Main fasta Swiss-Prot & TrEMBL (« canonical»)	Additional fasta Swiss-Prot & TrEMBL (« isoforms»)	Swiss-Prot & TrEMBL (canonical + isoforms)
Total	22 281*	36 800#	**59 081**

*UP000000589_10090.fasta.gz

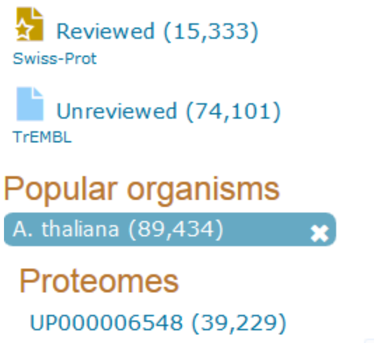
#UP000000589_10090_additional.fasta.gz

UniProt release 2017_05

* # ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/README

SIB Swiss Institute of Bioinformatics

Query + Download: ARATH



Reviewed (15,333)
Swiss-Prot

Unreviewed (74,101)
TrEMBL

Popular organisms

A. thaliana (89,434)

Proteomes

UP000006548 (39,229)

ARATH (TaxID: 3702) WEB	Swiss-Prot (canonical)	Swiss-Prot (isoforms)	TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (canonical + isoforms)	Swiss-Prot & TrEMBL (without Swiss-Prot isoforms)
Total	15 333	2 122	74 101	91 546	89 434
Proteome	15 260	2 122	23 969	**41 341**	39 229

ARATH FTP (README) Reference proteome	Main fasta Swiss-Prot & TrEMBL (« canonical»)	Additional fasta Swiss-Prot & TrEMBL (« isoforms)»	Swiss-Prot & TrEMBL (canonical + isoforms)
Total	27 510*	13 831#	**41 341**

*UP000006548_3702.fasta.gz

#UP000006548_3702_additional.fasta.gz

UniProt release 2017_05

* # ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/README

Swiss Institute of Bioinformatics

UniProt website and tools

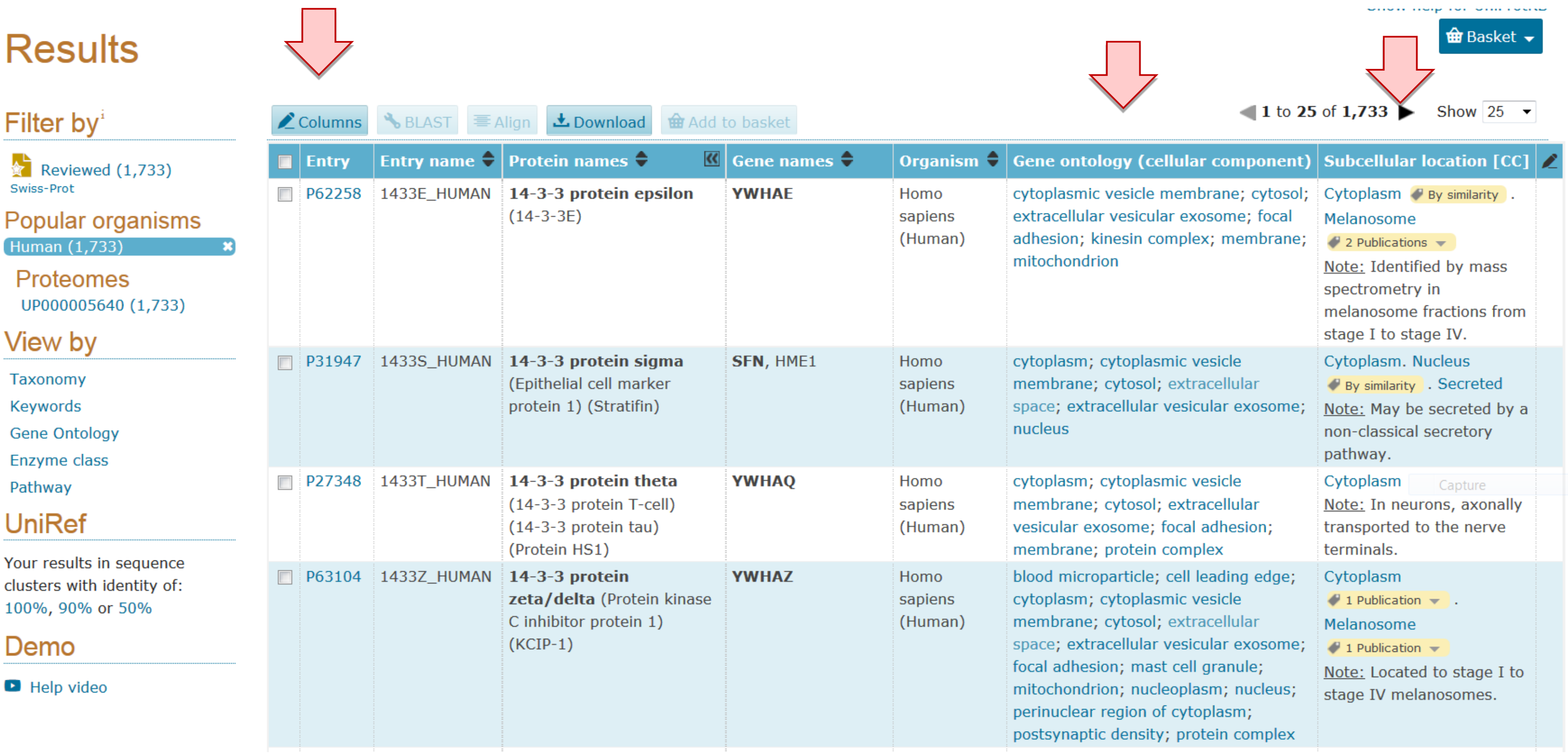
Swiss Institute of Bioinformatics

The UniProt web site – www.uniprot.org

- Powerful search engine, google-like and easy-to-use, but also supports very directed field searches
- Entry views, search result views and downloads are customizable
- The URL of a result page reflects the query; all pages and queries are **bookmarkable**, supporting programmatic access

- Tools: Blast, Align, Retrieve/Idmapping, Peptide search

Result pages: highly customizable (also available for Blast)



Results

Filter by

Reviewed (1,733)
Swiss-Prot

Popular organisms

Human (1,733) ✖

Proteomes

UP000005640 (1,733)

View by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

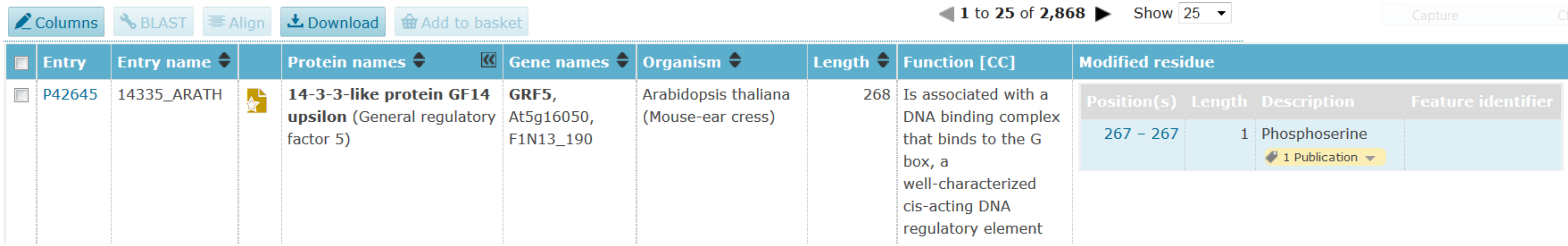
Your results in sequence clusters with identity of: 100%, 90% or 50%

Demo

Help video

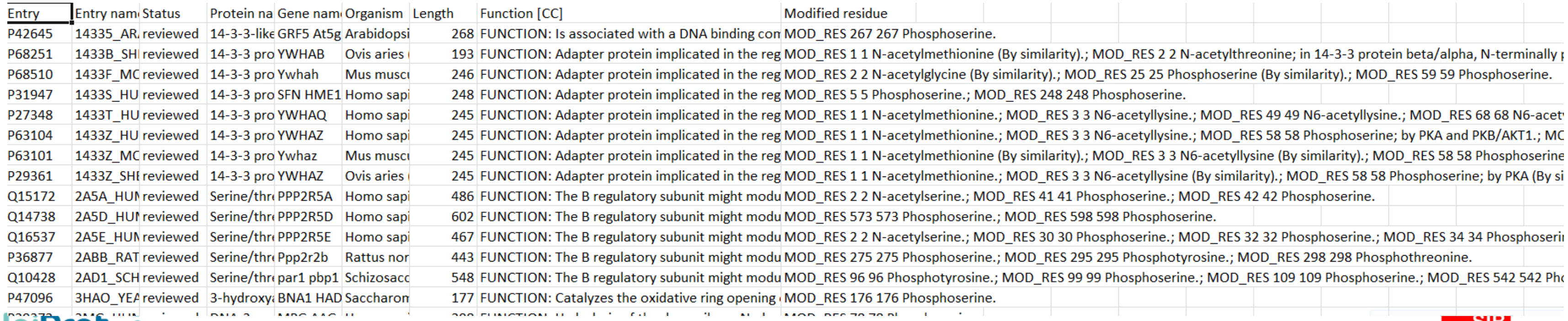
□	Entry	Entry name	Protein names	Gene names	Organism	Gene ontology (cellular component)	Subcellular location [CC]
□	P62258	1433E_HUMAN	**14-3-3 protein epsilon** (14-3-3E)	**YWHAE**	Homo sapiens (Human)	cytoplasmic vesicle membrane; cytosol; extracellular vesicular exosome; focal adhesion; kinesin complex; membrane; mitochondrion	Cytoplasm By similarity . Melanosome 2 Publications Note: Identified by mass spectrometry in melanosome fractions from stage I to stage IV.
□	P31947	1433S_HUMAN	**14-3-3 protein sigma** (Epithelial cell marker protein 1) (Stratifin)	**SFN**, HME1	Homo sapiens (Human)	cytoplasm; cytoplasmic vesicle membrane; cytosol; extracellular space; extracellular vesicular exosome; nucleus	Cytoplasm. Nucleus By similarity . Secreted Note: May be secreted by a non-classical secretory pathway.
□	P27348	1433T_HUMAN	**14-3-3 protein theta** (14-3-3 protein T-cell) (14-3-3 protein tau) (Protein HS1)	**YWHAQ**	Homo sapiens (Human)	cytoplasm; cytoplasmic vesicle membrane; cytosol; extracellular vesicular exosome; focal adhesion; membrane; protein complex	Cytoplasm Note: In neurons, axonally transported to the nerve terminals.
□	P63104	1433Z_HUMAN	**14-3-3 protein zeta/delta** (Protein kinase C inhibitor protein 1) (KCIP-1)	**YWHAZ**	Homo sapiens (Human)	blood microparticle; cell leading edge; cytoplasm; cytoplasmic vesicle membrane; cytosol; extracellular space; extracellular vesicular exosome; focal adhesion; mast cell granule; mitochondrion; nucleoplasm; nucleus; perinuclear region of cytoplasm; postsynaptic density; protein complex	Cytoplasm 1 Publication . Melanosome 1 Publication Note: Located to stage I to stage IV melanosomes.

Result page: download



☐	Entry	Entry name		Protein names	Gene names	Organism	Length	Function [CC]	Modified residue
☐	P42645	14335_ARATH		**14-3-3-like protein GF14 upsilon** (General regulatory factor 5)	**GRF5**, At5g16050, F1N13_190	Arabidopsis thaliana (Mouse-ear cress)	268	Is associated with a DNA binding complex that binds to the G box, a well-characterized cis-acting DNA regulatory element	Position(s): 267 – 267 Length: 1 Description: Phosphoserine 1 Publication Feature identifier:

Different formats (fasta, txt, excell, RDF, etc.)

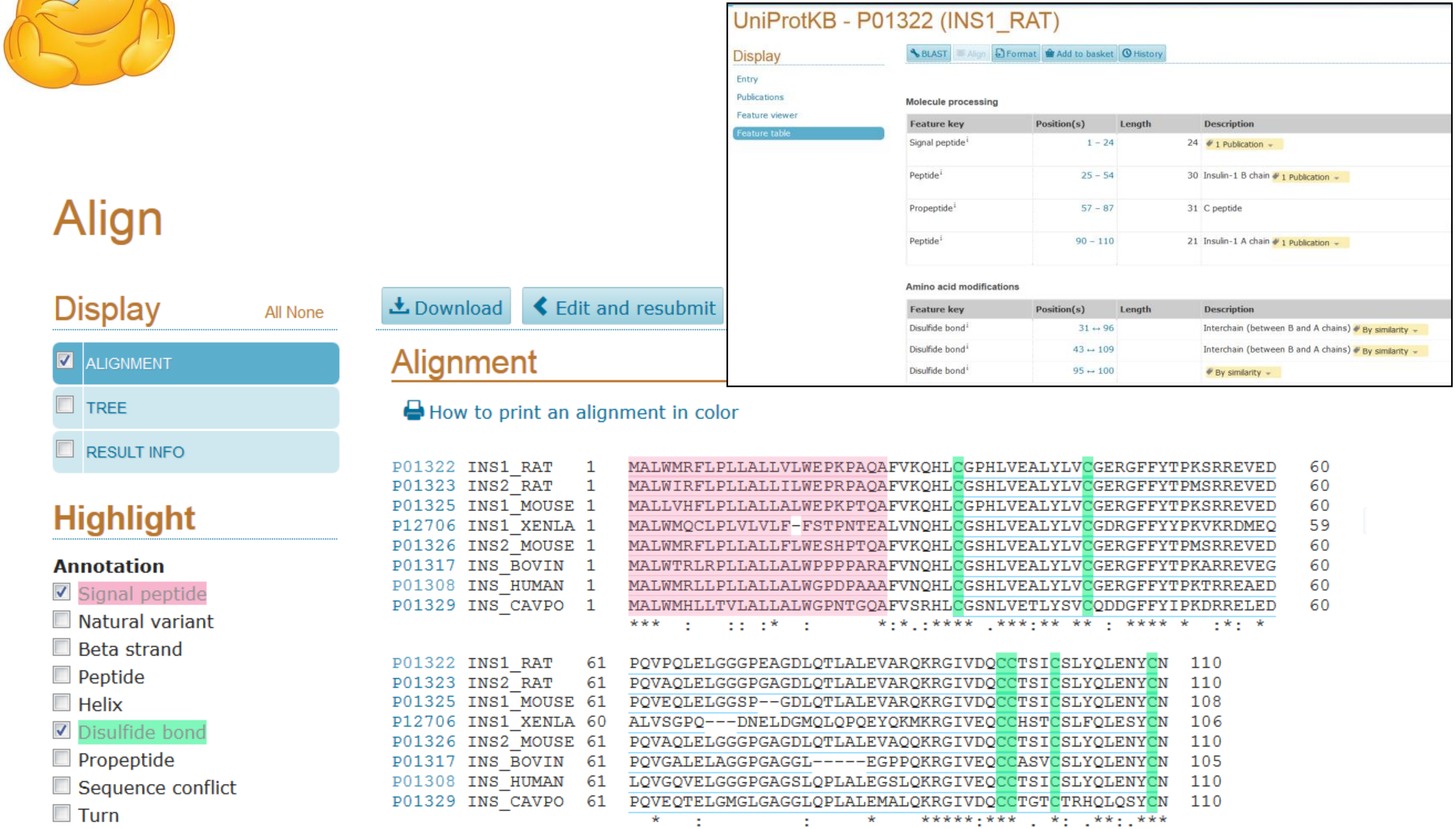


Entry	Entry nam	Status	Protein na	Gene nam	Organism	Length	Function [CC]	Modified residue
P42645	14335_AR.	reviewed	14-3-3-like	GRF5 At5g	Arabidopsi	268	FUNCTION: Is associated with a DNA binding con	MOD_RES 267 267 Phosphoserine.
P68251	1433B_SHI	reviewed	14-3-3 pro	YWHAB	Ovis aries (	193	FUNCTION: Adapter protein implicated in the reg	MOD_RES 1 1 N-acetylmethionine (By similarity).; MOD_RES 2 2 N-acetylthreonine; in 14-3-3 protein beta/alpha, N-terminally
P68510	1433F_MC	reviewed	14-3-3 pro	Ywhah	Mus muscu	246	FUNCTION: Adapter protein implicated in the reg	MOD_RES 2 2 N-acetylglycine (By similarity).; MOD_RES 25 25 Phosphoserine (By similarity).; MOD_RES 59 59 Phosphoserine.
P31947	1433S_HU	reviewed	14-3-3 pro	SFN HME1	Homo sapi	248	FUNCTION: Adapter protein implicated in the reg	MOD_RES 5 5 Phosphoserine.; MOD_RES 248 248 Phosphoserine.
P27348	1433T_HU	reviewed	14-3-3 pro	YWHAQ	Homo sapi	245	FUNCTION: Adapter protein implicated in the reg	MOD_RES 1 1 N-acetylmethionine.; MOD_RES 3 3 N6-acetyllysine.; MOD_RES 49 49 N6-acetyllysine.; MOD_RES 68 68 N6-acet
P63104	1433Z_HU	reviewed	14-3-3 pro	YWHAZ	Homo sapi	245	FUNCTION: Adapter protein implicated in the reg	MOD_RES 1 1 N-acetylmethionine.; MOD_RES 3 3 N6-acetyllysine.; MOD_RES 58 58 Phosphoserine; by PKA and PKB/AKT1.; MC
P63101	1433Z_MC	reviewed	14-3-3 pro	Ywhaz	Mus muscu	245	FUNCTION: Adapter protein implicated in the reg	MOD_RES 1 1 N-acetylmethionine (By similarity).; MOD_RES 3 3 N6-acetyllysine (By similarity).; MOD_RES 58 58 Phosphoserine
P29361	1433Z_SHI	reviewed	14-3-3 pro	YWHAZ	Ovis aries (	245	FUNCTION: Adapter protein implicated in the reg	MOD_RES 1 1 N-acetylmethionine.; MOD_RES 3 3 N6-acetyllysine (By similarity).; MOD_RES 58 58 Phosphoserine; by PKA (By si
Q15172	2A5A_HUI	reviewed	Serine/thr	PPP2R5A	Homo sapi	486	FUNCTION: The B regulatory subunit might modu	MOD_RES 2 2 N-acetylserine.; MOD_RES 41 41 Phosphoserine.; MOD_RES 42 42 Phosphoserine.
Q14738	2A5D_HUI	reviewed	Serine/thr	PPP2R5D	Homo sapi	602	FUNCTION: The B regulatory subunit might modu	MOD_RES 573 573 Phosphoserine.; MOD_RES 598 598 Phosphoserine.
Q16537	2A5E_HUI	reviewed	Serine/thr	PPP2R5E	Homo sapi	467	FUNCTION: The B regulatory subunit might modu	MOD_RES 2 2 N-acetylserine.; MOD_RES 30 30 Phosphoserine.; MOD_RES 32 32 Phosphoserine.; MOD_RES 34 34 Phosphoseri
P36877	2ABB_RAT	reviewed	Serine/thr	Ppp2r2b	Rattus nor	443	FUNCTION: The B regulatory subunit might modu	MOD_RES 275 275 Phosphoserine.; MOD_RES 295 295 Phosphotyrosine.; MOD_RES 298 298 Phosphothreonine.
Q10428	2AD1_SCH	reviewed	Serine/thr	par1 pbp1	Schizosacc	548	FUNCTION: The B regulatory subunit might modu	MOD_RES 96 96 Phosphotyrosine.; MOD_RES 99 99 Phosphoserine.; MOD_RES 109 109 Phosphoserine.; MOD_RES 542 542 Pho
P47096	3HAO_YEA	reviewed	3-hydroxy	BNA1 HAD	Saccharom	177	FUNCTION: Catalyzes the oxidative ring opening	MOD_RES 176 176 Phosphoserine.

UniProt

SIB Swiss Institute of Bioinformatics

Highlight sequence annotation in alignment (BLAST or multiple alignment)



UniProtKB - P01322 (INS1_RAT)

Display

- Entry
- Publications
- Feature viewer
- Feature table

Molecule processing

Feature key	Position(s)	Length	Description
Signal peptide	1 – 24	24	1 Publication
Peptide	25 – 54	30	Insulin-1 B chain 1 Publication
Propeptide	57 – 87	31	C peptide
Peptide	90 – 110	21	Insulin-1 A chain 1 Publication

Amino acid modifications

Feature key	Position(s)	Length	Description
Disulfide bond	31 ↔ 96		Interchain (between B and A chains) By similarity
Disulfide bond	43 ↔ 109		Interchain (between B and A chains) By similarity
Disulfide bond	95 ↔ 100		By similarity

Align

Display All None

- ☑ ALIGNMENT
- □ TREE
- □ RESULT INFO

Highlight

Annotation

- ☑ Signal peptide
- □ Natural variant
- □ Beta strand
- □ Peptide
- □ Helix
- ☑ Disulfide bond
- □ Propeptide
- □ Sequence conflict
- □ Turn

Alignment

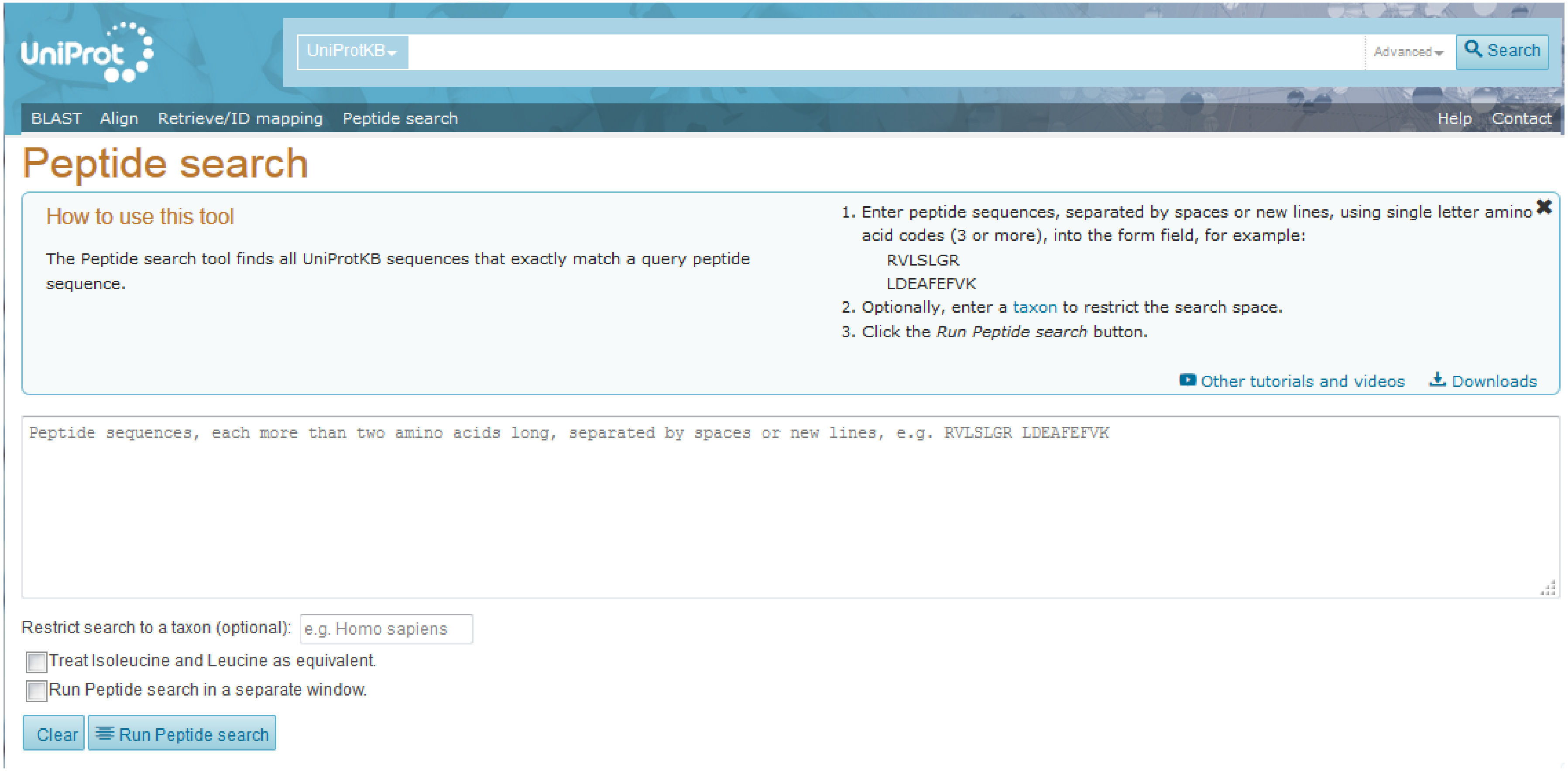
How to print an alignment in color

```
P01322 INS1_RAT   1    MALWMRFLPLLALLVLWEPKPAQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED   60
P01323 INS2_RAT   1    MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVED   60
P01325 INS1_MOUSE 1    MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED   60
P12706 INS1_XENLA 1    MALWMQCLPLVLVLF-FSTPNTEALVNQHLCGSHLVEALYLVCGDRGFFYYPKVKRDMEQ   59
P01326 INS2_MOUSE 1    MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVED   60
P01317 INS_BOVIN  1    MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG   60
P01308 INS_HUMAN  1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
P01329 INS_CAVPO  1    MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELED   60
                       ***  :   :: :*  :      *:*.:**** .***:** ** : **** *  :*: *

P01322 INS1_RAT   61   PQVPQLELGGGPEAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN  110
P01323 INS2_RAT   61   PQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN  110
P01325 INS1_MOUSE 61   PQVEQLELGGSP--GDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN  108
P12706 INS1_XENLA 60   ALVSGPQ---DNELDGMQLQPQEYQKMKRGIVEQCCHSTCSLFQLESYCN  106
P01326 INS2_MOUSE 61   PQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN  110
P01317 INS_BOVIN  61   PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN  105
P01308 INS_HUMAN  61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
P01329 INS_CAVPO  61   PQVEQTELGMGLGAGGLQPLALEMALQKRGIVDQCCTGTCTRHQLQSYCN  110
                         *   :         :     *    *****:*** . *: .**:.***
```

UniProt

SIB Swiss Institute of Bioinformatics

Peptide search



Peptide search

How to use this tool

The Peptide search tool finds all UniProtKB sequences that exactly match a query peptide sequence.

1. Enter peptide sequences, separated by spaces or new lines, using single letter amino acid codes (3 or more), into the form field, for example:
   RVLSLGR
   LDEAFEFVK
2. Optionally, enter a taxon to restrict the search space.
3. Click the *Run Peptide search* button.

Other tutorials and videos Downloads

Peptide sequences, each more than two amino acids long, separated by spaces or new lines, e.g. RVLSLGR LDEAFEFVK

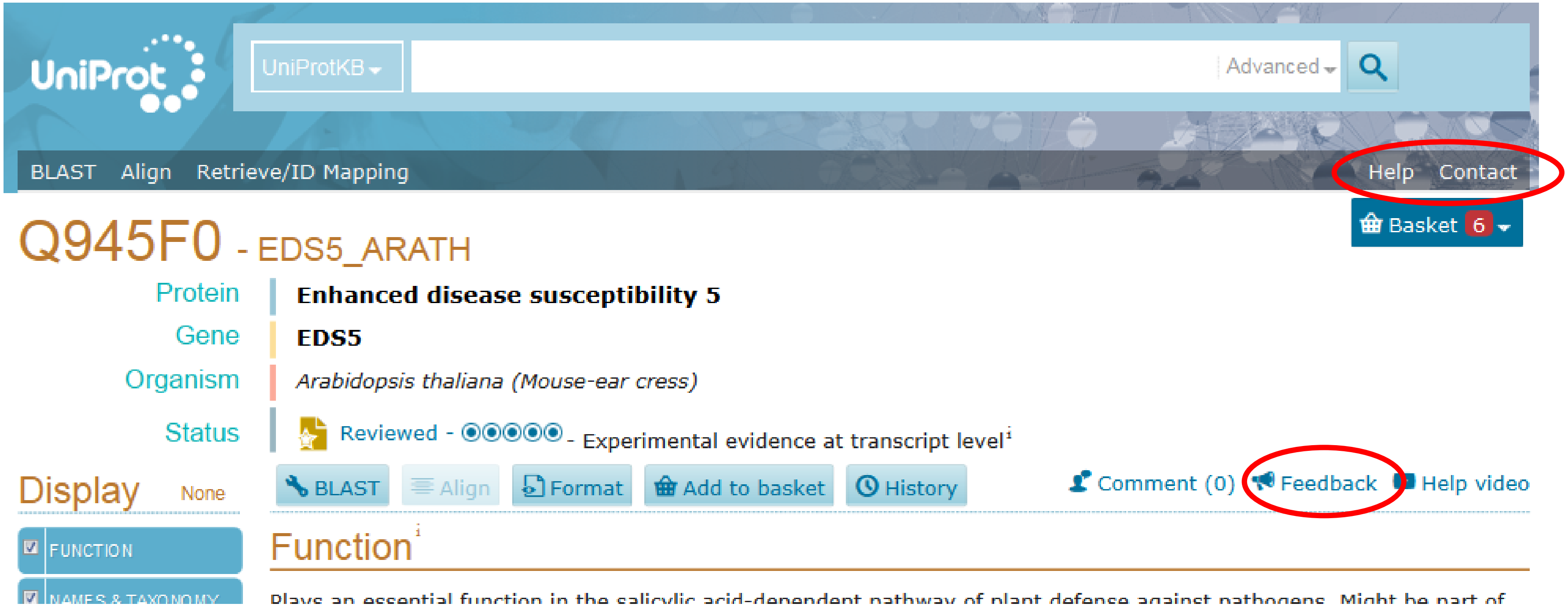
Restrict search to a taxon (optional): e.g. Homo sapiens

□ Treat Isoleucine and Leucine as equivalent.

□ Run Peptide search in a separate window.

http://www.uniprot.org/peptidesearch/

help@uniprot.org



Q945F0 - EDS5_ARATH

Protein: Enhanced disease susceptibility 5

Gene: EDS5

Organism: *Arabidopsis thaliana (Mouse-ear cress)*

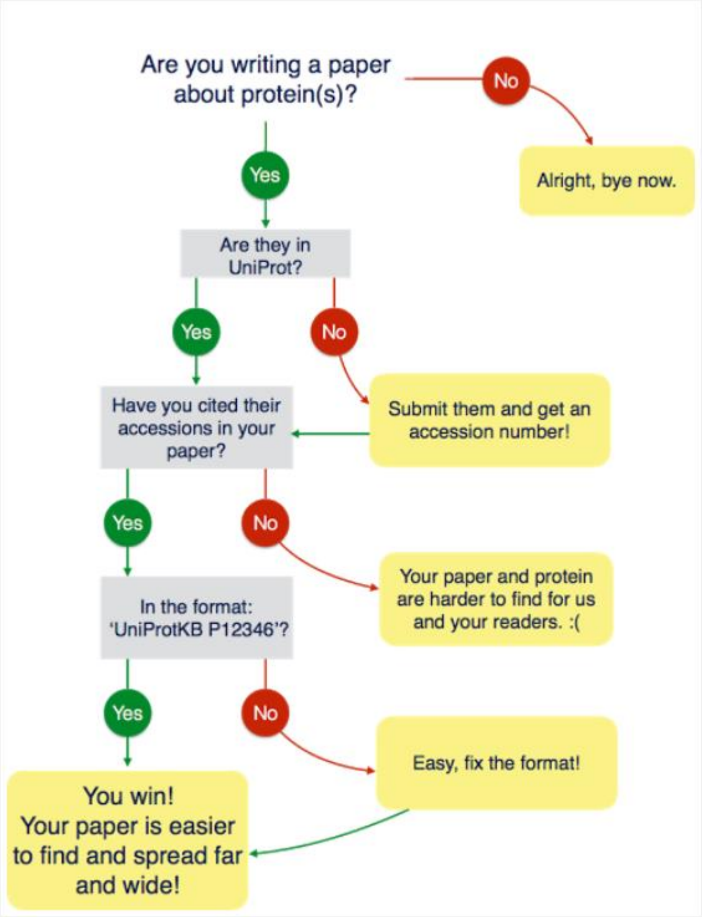
Status: Reviewed - ●●●●● - Experimental evidence at transcript level[i]

Function[i]

Plays an essential function in the salicylic acid-dependent pathway of plant defense against pathogens. Might be part of

Swiss Institute of Bioinformatics

How can you increase the impact of your research papers and contribute to UniProt?



Are you writing a paper about protein(s)?

No

Alright, bye now.

Yes

Are they in UniProt?

Yes

No

Submit them and get an accession number!

Have you cited their accessions in your paper?

Yes

No

Your paper and protein are harder to find for us and your readers. :(

In the format: 'UniProtKB P12346'?

Yes

No

Easy, fix the format!

You win! Your paper is easier to find and spread far and wide!

http://insideuniprot.blogspot.ch/2016/09/how-can-you-increase-impact-of-your.html

UniProt

SIB Swiss Institute of Bioinformatics

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

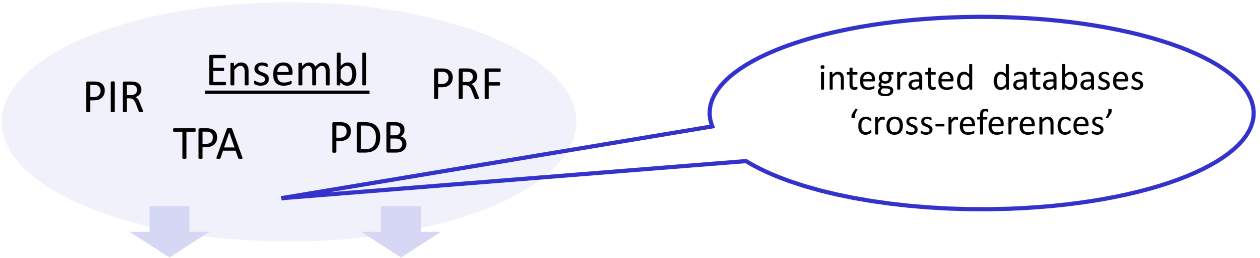
UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

➡ **NCBI protein (RefSeq)**

Practicals

Major ‘general’ protein sequence database ‘sources’


PIR Ensembl PRF

TPA PDB

integrated databases
‘cross-references’

UniProtKB: **Swiss-Prot + TrEMBL**

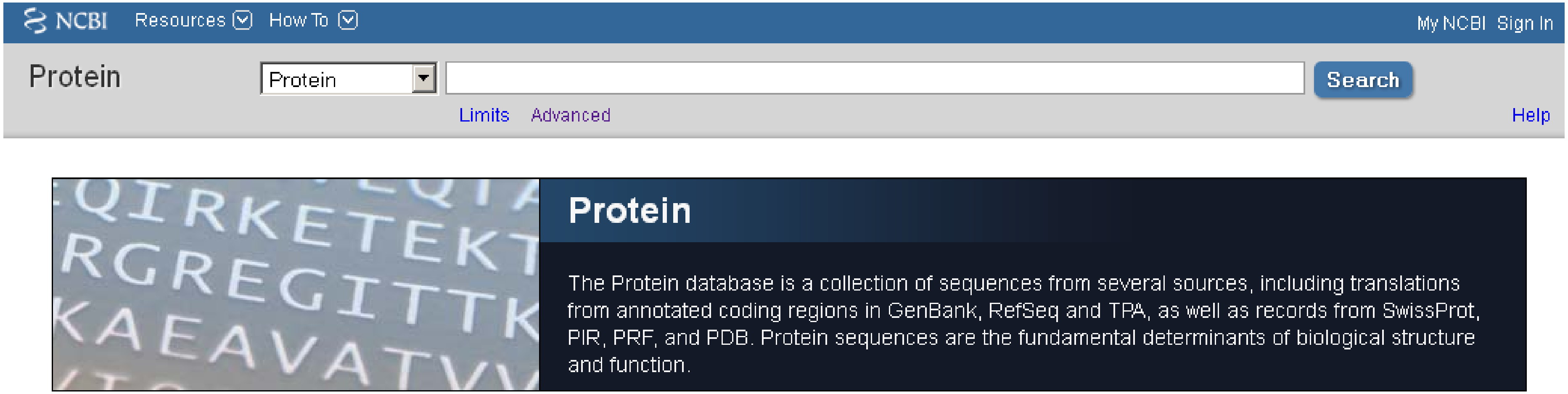
databases are kept
separated

NCBI-nr: **Swiss-Prot + TrEMBL** + **GenPept** + PIR + PDB + PRF + **RefSeq** + TPA

not complete !!!
(only entries created before 2007 ?)

Ensembl and Refseq: gene prediction

NCBI nr - Entrez ‘protein’



Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

Using Protein	Protein Tools	Other Resources
Quick Start Guide	BLAST	GenBank Home
FAQ	LinkOut	RefSeq Home
Help	E-Utilities	CDD
GenBank FTP	Blink	Structure
RefSeq FTP	Batch Entrez	

NCBI-nr

- GenPept (source: GenBank; translated CDS)
- RefSeq
- TPA (third part annotation)

- Swiss-Prot (does not include isoform sequences)
- PIR (not updated since 2003)
- PRF (journal scan of ‘published’ peptide)
- PDB (Protein Data Bank, 3D structure)
- TrEMBL (some entries….)

NCBI-nr: Swiss-Prot + GenPept + PIR + PDB + PRF + RefSeq + TPA

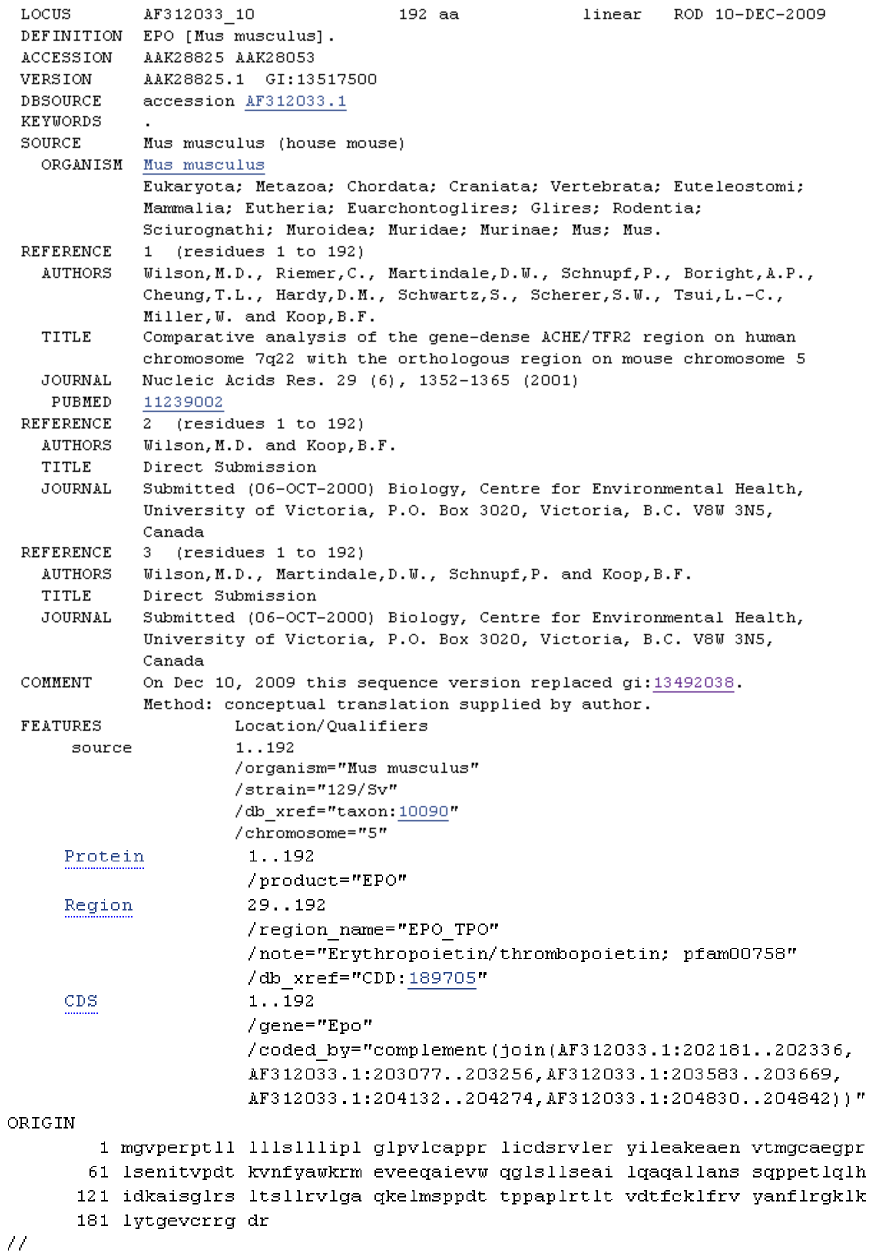
GenPept

Translation from annotated CDS in GenBank

Contains all translated CDS annotated in
GenBank/EMBL/DDBJ sequences

- equivalent to UniProtKB/TrEMBL,
except that it is
redundant with other databases
(Swiss-Prot, RefSeq, PIR….)

GenPept: ‘translations from all annotated coding regions (CDS) in GenBank’



```
LOCUS       AF312033_10              192 aa            linear   ROD 10-DEC-2009
DEFINITION  EPO [Mus musculus].
ACCESSION   AAK28825 AAK28053
VERSION     AAK28825.1  GI:13517500
DBSOURCE    accession AF312033.1
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (residues 1 to 192)
  AUTHORS   Wilson,M.D., Riemer,C., Martindale,D.W., Schnupf,P., Boright,A.P.,
            Cheung,T.L., Hardy,D.M., Schwartz,S., Scherer,S.W., Tsui,L.-C.,
            Miller,W. and Koop,B.F.
  TITLE     Comparative analysis of the gene-dense ACHE/TFR2 region on human
            chromosome 7q22 with the orthologous region on mouse chromosome 5
  JOURNAL   Nucleic Acids Res. 29 (6), 1352-1365 (2001)
   PUBMED   11239002
REFERENCE   2  (residues 1 to 192)
  AUTHORS   Wilson,M.D. and Koop,B.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-OCT-2000) Biology, Centre for Environmental Health,
            University of Victoria, P.O. Box 3020, Victoria, B.C. V8W 3N5,
            Canada
REFERENCE   3  (residues 1 to 192)
  AUTHORS   Wilson,M.D., Martindale,D.W., Schnupf,P. and Koop,B.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-OCT-2000) Biology, Centre for Environmental Health,
            University of Victoria, P.O. Box 3020, Victoria, B.C. V8W 3N5,
            Canada
COMMENT     On Dec 10, 2009 this sequence version replaced gi:13492038.
            Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source          1..192
                     /organism="Mus musculus"
                     /strain="129/Sv"
                     /db_xref="taxon:10090"
                     /chromosome="5"
     Protein          1..192
                      /product="EPO"
     Region           29..192
                      /region_name="EPO_TPO"
                      /note="Erythropoietin/thrombopoietin; pfam00758"
                      /db_xref="CDD:189705"
     CDS              1..192
                      /gene="Epo"
                      /coded_by="complement(join(AF312033.1:202181..202336,
                      AF312033.1:203077..203256,AF312033.1:203583..203669,
                      AF312033.1:204132..204274,AF312033.1:204830..204842))"
ORIGIN
        1 mgvperptll lllslllipl glpvlcappr licdsrvler yileakeaen vtmgcaegpr
       61 lsenitvpdt kvnfyawkrm eveeqaievw qglsllseai lqaqallans sqppetlqlh
      121 idkaisglrs ltsllrvlga qkelmsppdt tppaplrtlt vdtfcklfrv yanflrgklk
      181 lytgevcrrg dr
//
```

Annotation according to the submitter

No GO term !

NCBI-nr: Swiss-Prot + GenPept + PIR + PDB + PRF + RefSeq + TPA

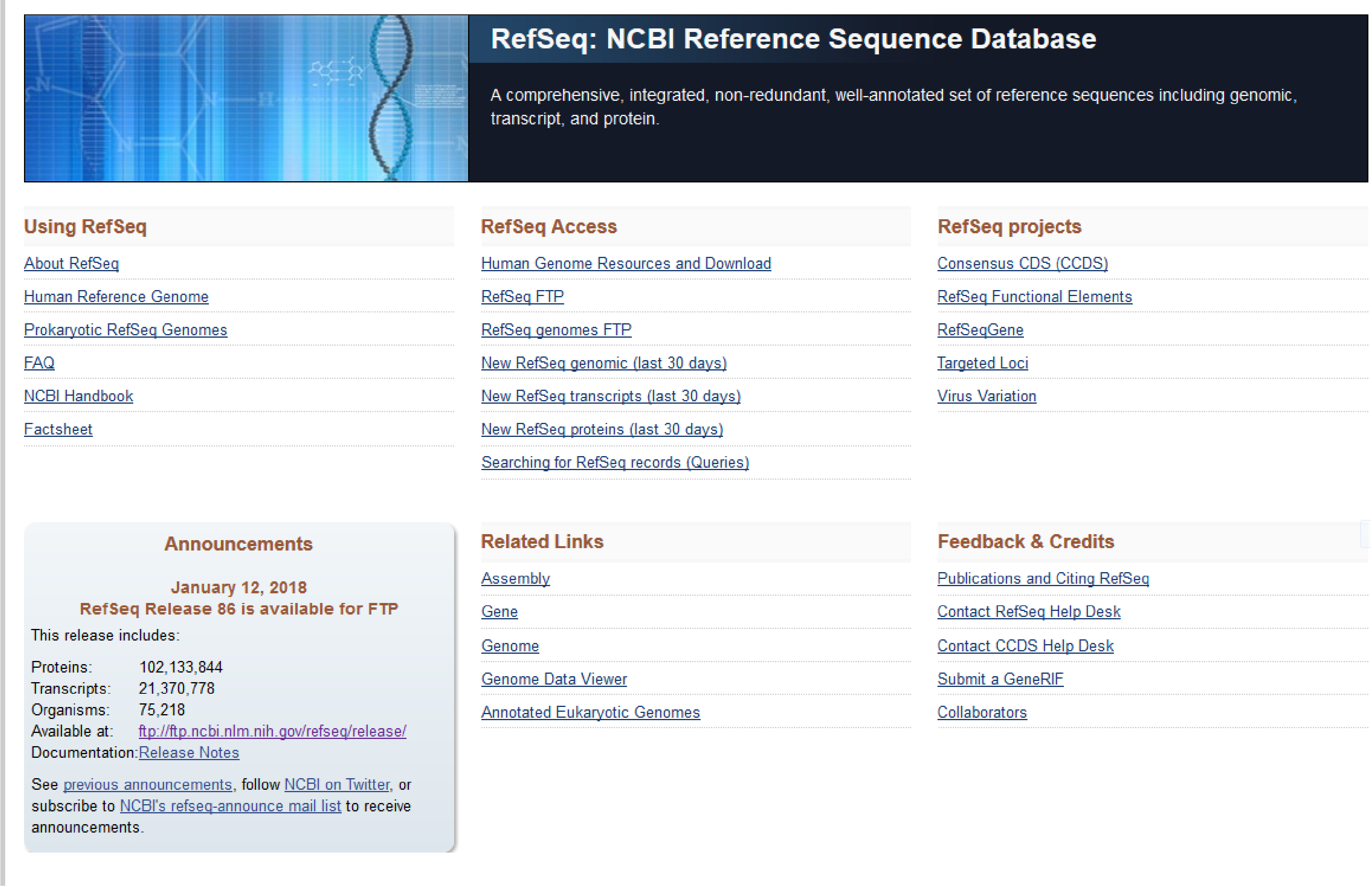
RefSeq

Produced by NCBI and NLM

http://www.ncbi.nlm.nih.gov/RefSeq/

http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch18.pdf

FAQ: http://www.ncbi.nlm.nih.gov/books/NBK50679/



RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Using RefSeq

- About RefSeq
- Human Reference Genome
- Prokaryotic RefSeq Genomes
- FAQ
- NCBI Handbook
- Factsheet

RefSeq Access

- Human Genome Resources and Download
- RefSeq FTP
- RefSeq genomes FTP
- New RefSeq genomic (last 30 days)
- New RefSeq transcripts (last 30 days)
- New RefSeq proteins (last 30 days)
- Searching for RefSeq records (Queries)

RefSeq projects

- Consensus CDS (CCDS)
- RefSeq Functional Elements
- RefSeqGene
- Targeted Loci
- Virus Variation

Announcements

January 12, 2018
RefSeq Release 86 is available for FTP

This release includes:

Proteins:	102,133,844
Transcripts:	21,370,778
Organisms:	75,218
Available at:	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/
Documentation:	Release Notes

See previous announcements, follow NCBI on Twitter, or subscribe to NCBI's refseq-announce mail list to receive announcements.

Related Links

- Assembly
- Gene
- Genome
- Genome Data Viewer
- Annotated Eukaryotic Genomes

Feedback & Credits

- Publications and Citing RefSeq
- Contact RefSeq Help Desk
- Contact CCDS Help Desk
- Submit a GeneRIF
- Collaborators

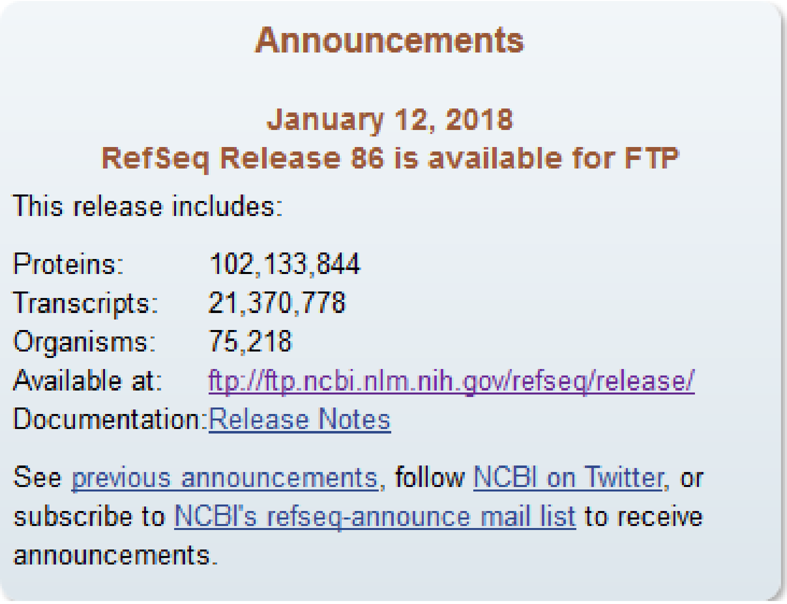
http://www.ncbi.nlm.nih.gov/refseq/

RefSeq

RefSeq: The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redondant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

one mRNA sequence -> one entry

Different entries for identical protein sequences



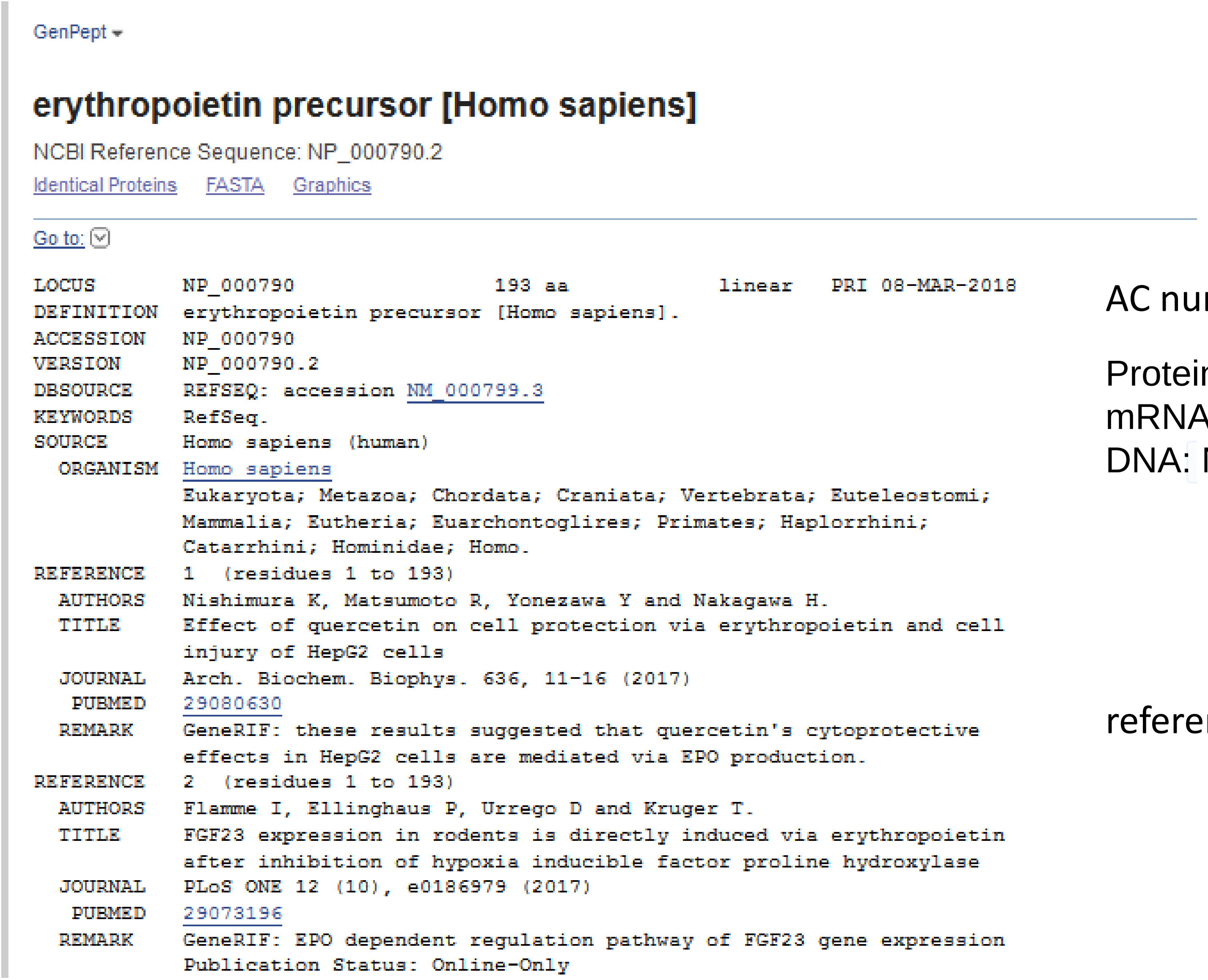
Announcements

January 12, 2018
RefSeq Release 86 is available for FTP

This release includes:

Proteins:	102,133,844
Transcripts:	21,370,778
Organisms:	75,218
Available at:	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/
Documentation:	Release Notes

See previous announcements, follow NCBI on Twitter, or subscribe to NCBI's refseq-announce mail list to receive announcements.



GenPept

erythropoietin precursor [Homo sapiens]

NCBI Reference Sequence: NP_000790.2

Identical Proteins   FASTA   Graphics

Go to:

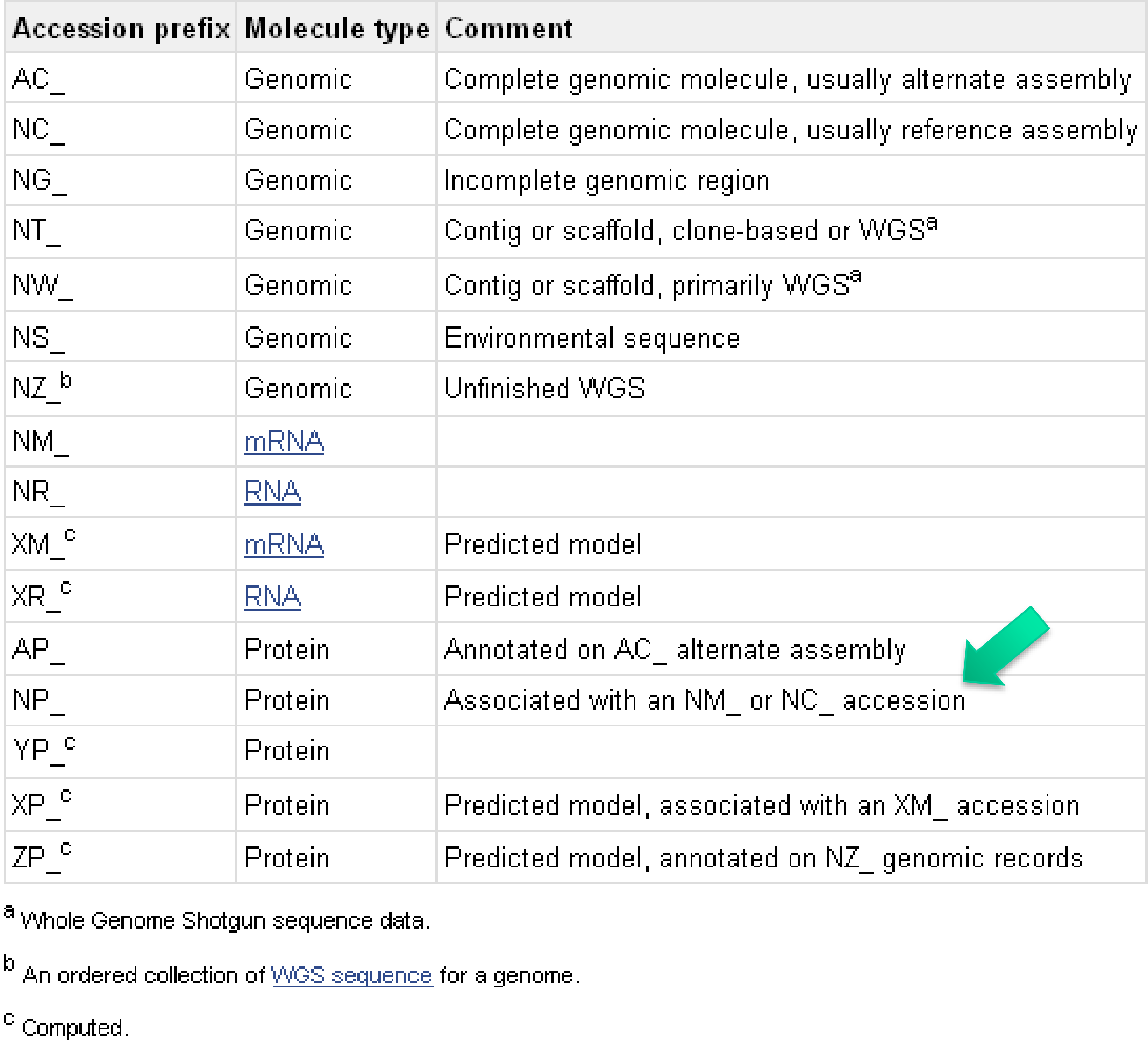
```
LOCUS       NP_000790                193 aa            linear   PRI 08-MAR-2018
DEFINITION  erythropoietin precursor [Homo sapiens].
ACCESSION   NP_000790
VERSION     NP_000790.2
DBSOURCE    REFSEQ: accession NM_000799.3
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 193)
  AUTHORS   Nishimura K, Matsumoto R, Yonezawa Y and Nakagawa H.
  TITLE     Effect of quercetin on cell protection via erythropoietin and cell
            injury of HepG2 cells
  JOURNAL   Arch. Biochem. Biophys. 636, 11-16 (2017)
   PUBMED   29080630
  REMARK    GeneRIF: these results suggested that quercetin's cytoprotective
            effects in HepG2 cells are mediated via EPO production.
REFERENCE   2  (residues 1 to 193)
  AUTHORS   Flamme I, Ellinghaus P, Urrego D and Kruger T.
  TITLE     FGF23 expression in rodents is directly induced via erythropoietin
            after inhibition of hypoxia inducible factor proline hydroxylase
  JOURNAL   PLoS ONE 12 (10), e0186979 (2017)
   PUBMED   29073196
  REMARK    GeneRIF: EPO dependent regulation pathway of FGF23 gene expression
            Publication Status: Online-Only
```

AC number

Protein: NP_
mRNA: NM_
DNA: NC_

references

Accession number

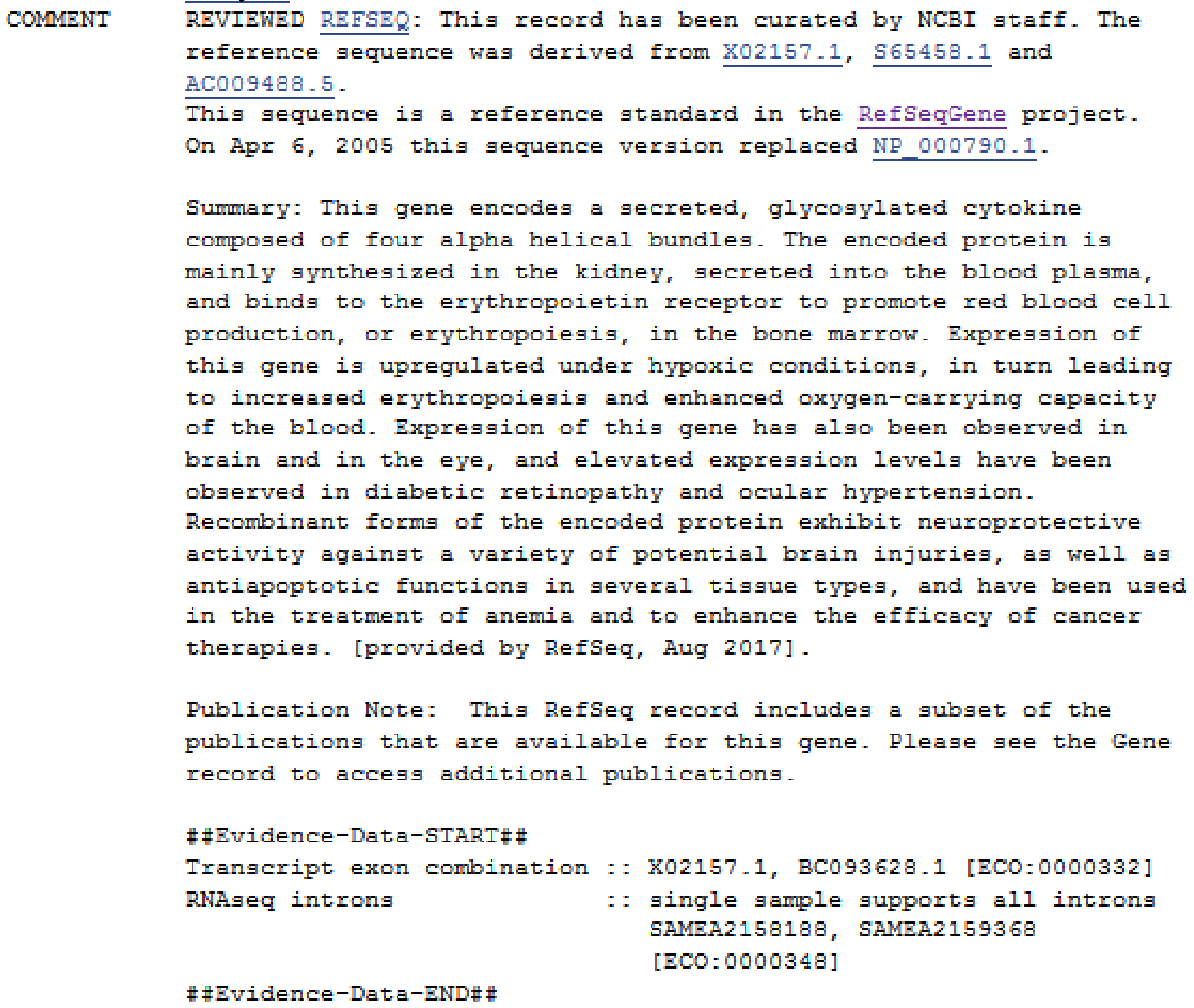


Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

[a] Whole Genome Shotgun sequence data.

[b] An ordered collection of WGS sequence for a genome.

[c] Computed.



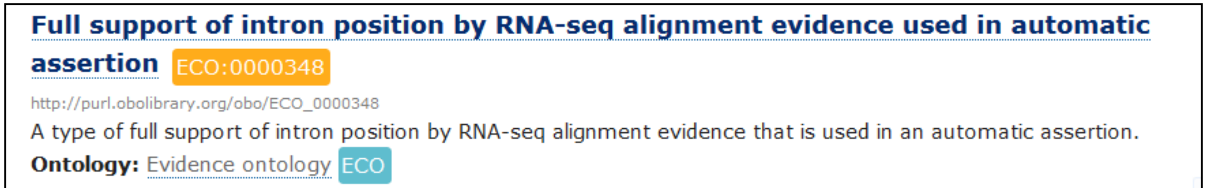
```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from X02157.1, S65458.1 and
            AC009488.5.
            This sequence is a reference standard in the RefSeqGene project.
            On Apr 6, 2005 this sequence version replaced NP_000790.1.

            Summary: This gene encodes a secreted, glycosylated cytokine
            composed of four alpha helical bundles. The encoded protein is
            mainly synthesized in the kidney, secreted into the blood plasma,
            and binds to the erythropoietin receptor to promote red blood cell
            production, or erythropoiesis, in the bone marrow. Expression of
            this gene is upregulated under hypoxic conditions, in turn leading
            to increased erythropoiesis and enhanced oxygen-carrying capacity
            of the blood. Expression of this gene has also been observed in
            brain and in the eye, and elevated expression levels have been
            observed in diabetic retinopathy and ocular hypertension.
            Recombinant forms of the encoded protein exhibit neuroprotective
            activity against a variety of potential brain injuries, as well as
            antiapoptotic functions in several tissue types, and have been used
            in the treatment of anemia and to enhance the efficacy of cancer
            therapies. [provided by RefSeq, Aug 2017].

            Publication Note:  This RefSeq record includes a subset of the
            publications that are available for this gene. Please see the Gene
            record to access additional publications.

            ##Evidence-Data-START##
            Transcript exon combination :: X02157.1, BC093628.1 [ECO:0000332]
            RNAseq introns              :: single sample supports all introns
                                           SAMEA2158188, SAMEA2159368
                                           [ECO:0000348]
            ##Evidence-Data-END##
```

Annotation

No GO term !

Sequence origin

+ ECO code



Full support of intron position by RNA-seq alignment evidence used in automatic assertion ECO:0000348

http://purl.obolibrary.org/obo/ECO_0000348

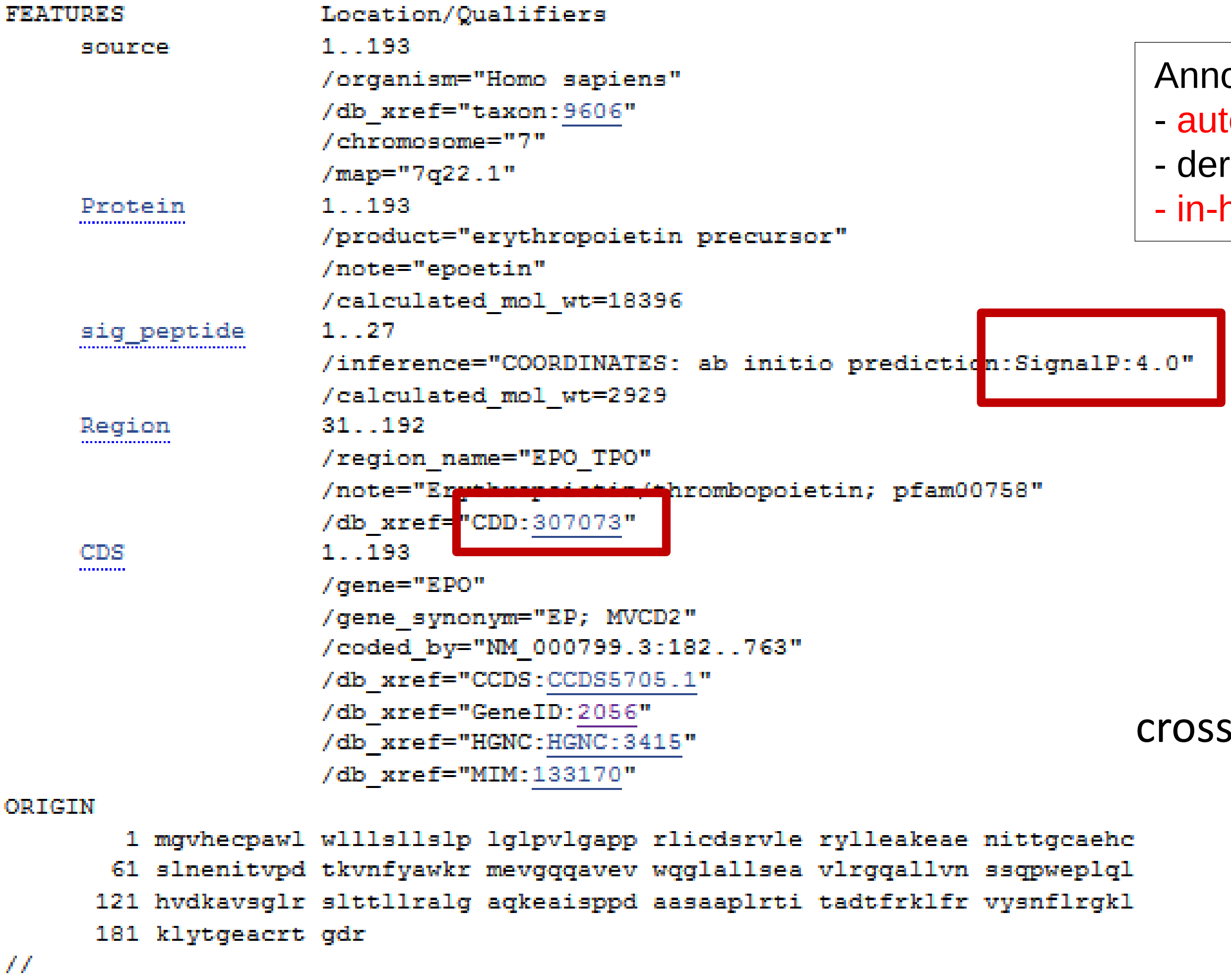
A type of full support of intron position by RNA-seq alignment evidence that is used in an automatic assertion.

Ontology: Evidence ontology ECO

RefSeq

	manual annotation
GENOME ANNOTATION	No
INFERRED	No
MODEL	No
PREDICTED	No
PROVISIONAL	No
REVIEWED	**Yes** (sequence + functional information and features)
VALIDATED	**Yes** (initial sequence)
Whole Genome Sequencing (WGS)	No

http://www.ncbi.nlm.nih.gov/RefSeq/



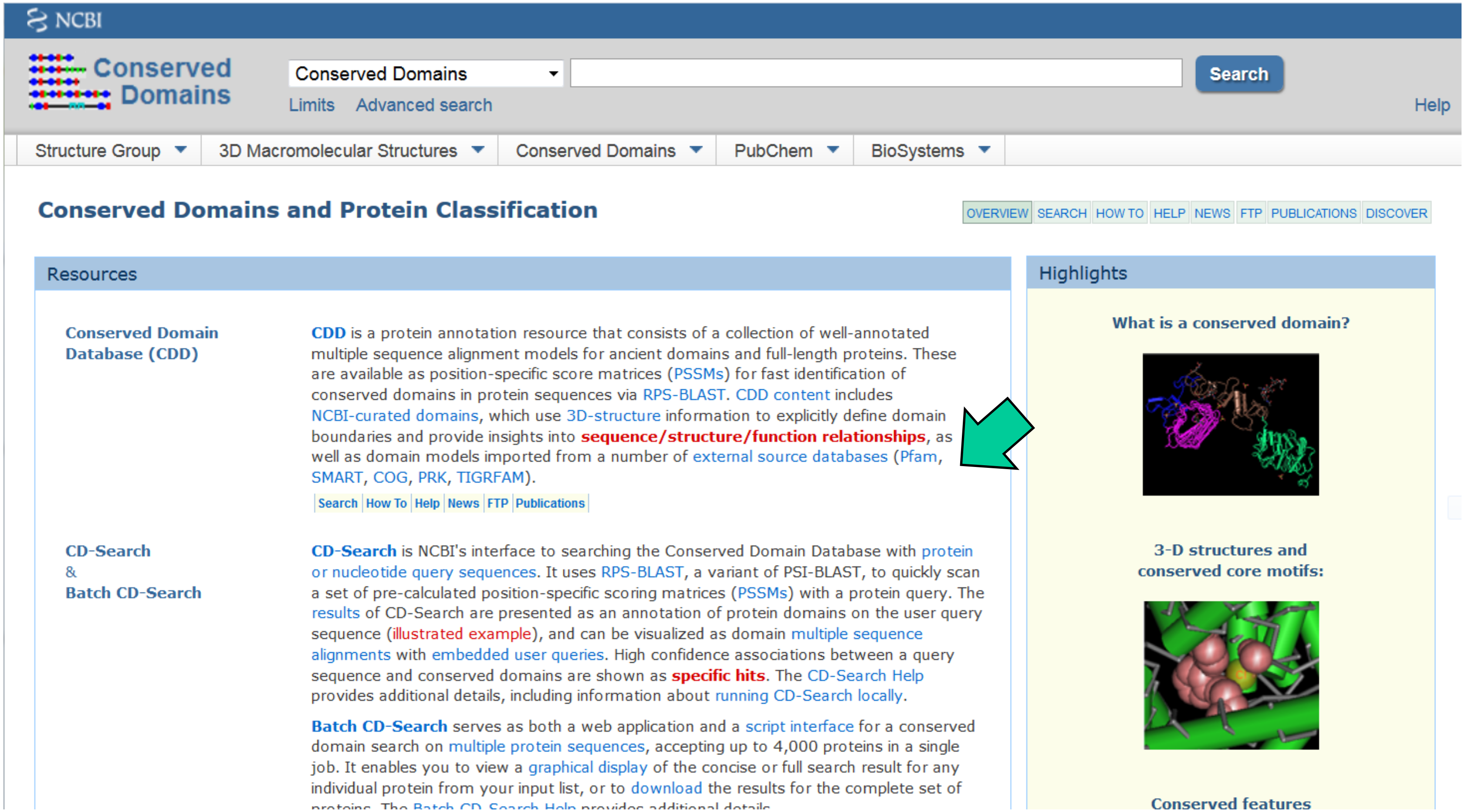
```
FEATURES             Location/Qualifiers
     source          1..193
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="7"
                     /map="7q22.1"
     Protein         1..193
                     /product="erythropoietin precursor"
                     /note="epoetin"
                     /calculated_mol_wt=18396
     sig_peptide     1..27
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
                     /calculated_mol_wt=2929
     Region          31..192
                     /region_name="EPO_TPO"
                     /note="Erythropoietin/thrombopoietin; pfam00758"
                     /db_xref="CDD:307073"
     CDS             1..193
                     /gene="EPO"
                     /gene_synonym="EP; MVCD2"
                     /coded_by="NM_000799.3:182..763"
                     /db_xref="CCDS:CCDS5705.1"
                     /db_xref="GeneID:2056"
                     /db_xref="HGNC:HGNC:3415"
                     /db_xref="MIM:133170"
ORIGIN
        1 mgvhecpawl wlllsllslp lglpvlgapp rlicdsrvle rylleakeae nittgcaehc
       61 slnenitvpd tkvnfyawkr mevgqqavev wqglallsea vlrgqallvn ssqpweplql
      121 hvdkavsglr slttllralg aqkeaisppd aasaaplrti tadtfrklfr vysnflrgkl
      181 klytgeacrt gdr
//
```

Annotation
- automated (CDD)
- derived from Swiss-Prot
- in-house (ab initio)

cross-references

sequence

Automated annotation: CDD



Conserved Domains and Protein Classification

OVERVIEW | SEARCH | HOW TO | HELP | NEWS | FTP | PUBLICATIONS | DISCOVER

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**. The CD-Search Help provides additional details, including information about running CD-Search locally.

Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 4,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

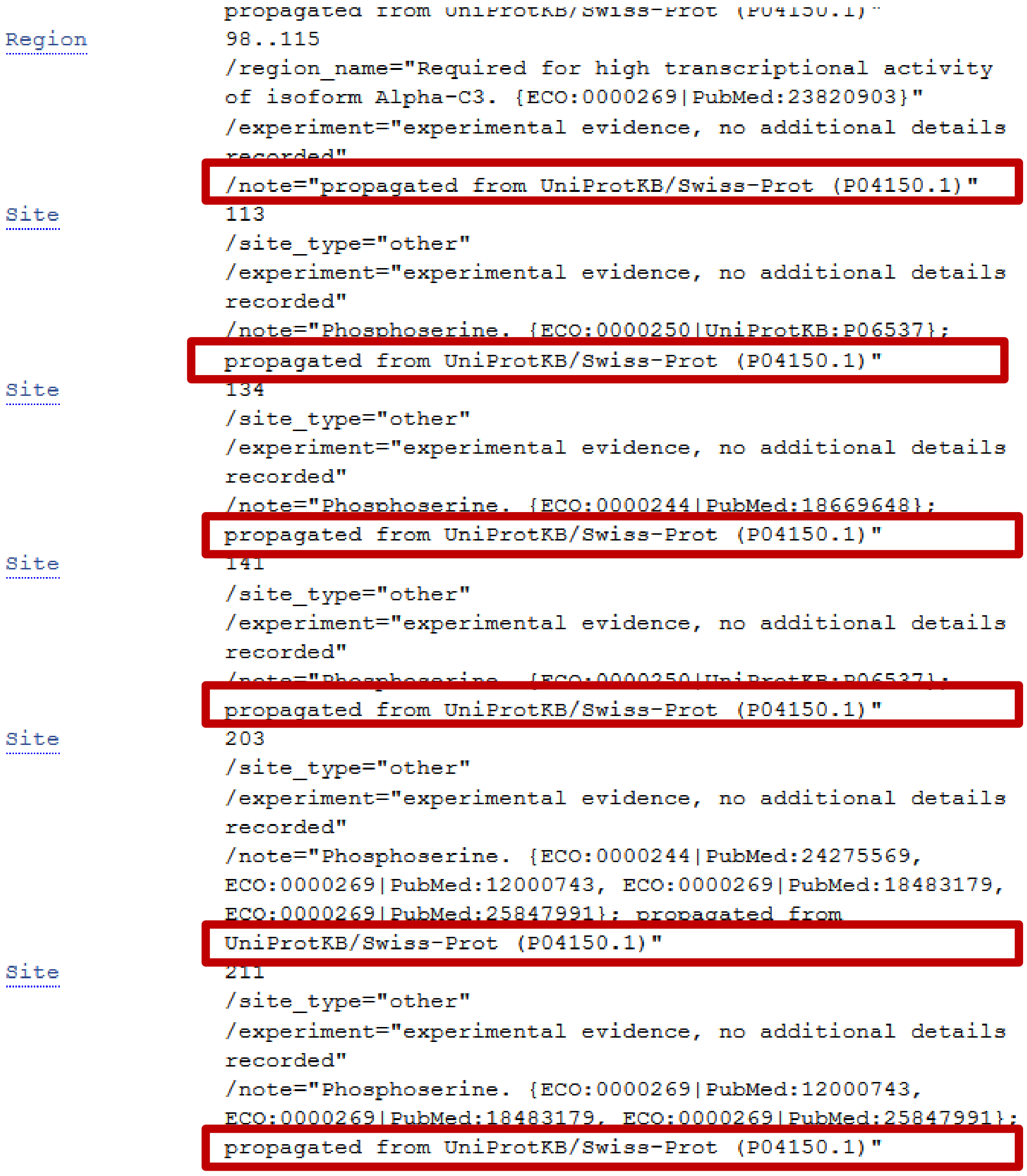
Highlights

What is a conserved domain?

3-D structures and conserved core motifs:

Conserved features

https://www.ncbi.nlm.nih.gov/protein/NP_000167.1

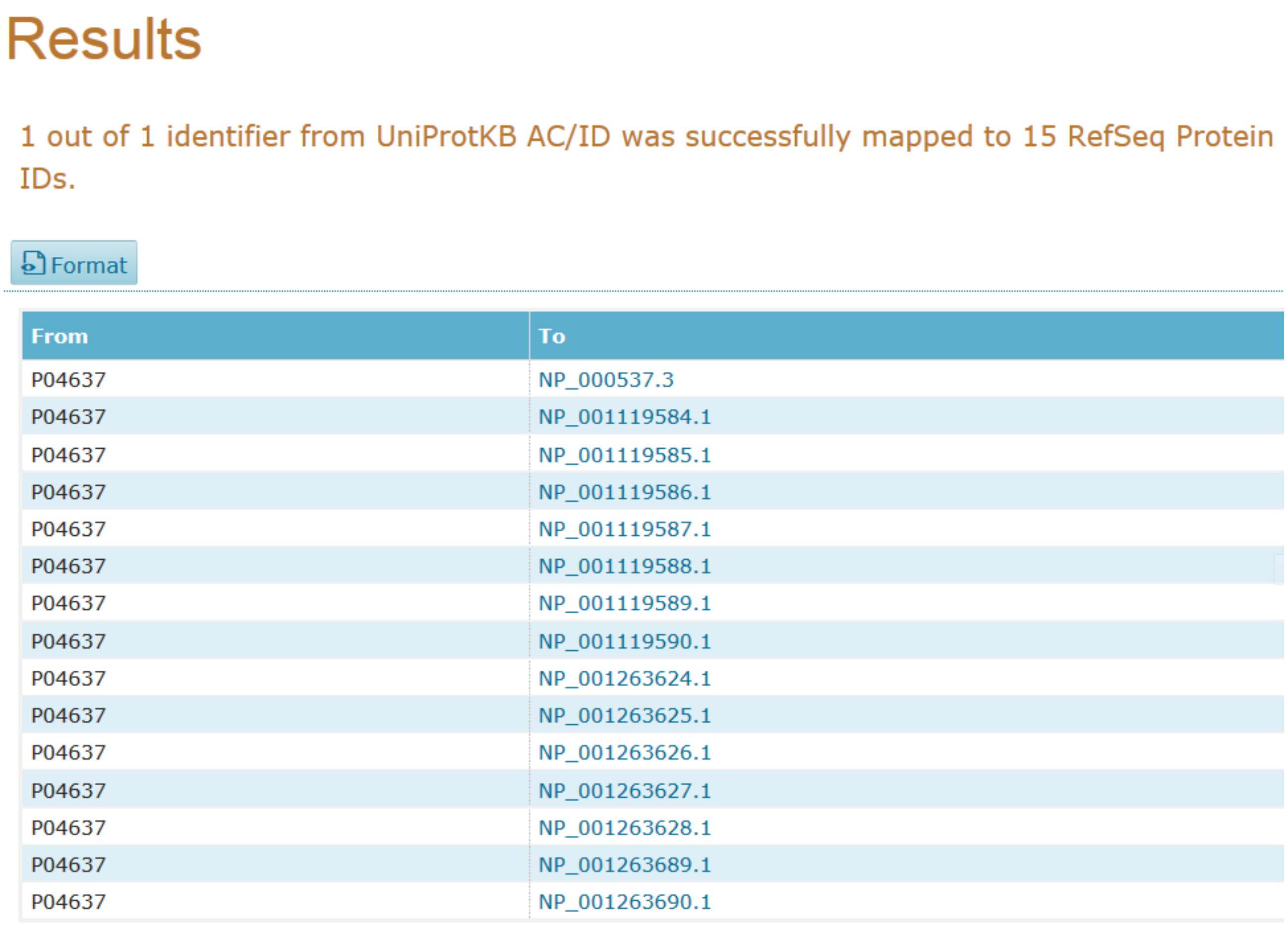


```
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Region          98..115
                     /region_name="Required for high transcriptional activity
                     of isoform Alpha-C3. {ECO:0000269|PubMed:23820903}"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            113
                     /site_type="other"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphoserine. {ECO:0000250|UniProtKB:P06537};
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            134
                     /site_type="other"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphoserine. {ECO:0000244|PubMed:18669648};
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            141
                     /site_type="other"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphoserine. {ECO:0000250|UniProtKB:P06537};
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
     Site            203
                     /site_type="other"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphoserine. {ECO:0000244|PubMed:24275569,
                     ECO:0000269|PubMed:12000743, ECO:0000269|PubMed:18483179,
                     ECO:0000269|PubMed:25847991}; propagated from
                     UniProtKB/Swiss-Prot (P04150.1)"
     Site            211
                     /site_type="other"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Phosphoserine. {ECO:0000269|PubMed:12000743,
                     ECO:0000269|PubMed:18483179, ECO:0000269|PubMed:25847991};
                     propagated from UniProtKB/Swiss-Prot (P04150.1)"
```

Annotation
- automated (CDD)
- derived from Swiss-Prot
- in-house (ab initio)

UniProtKB/Swiss-Prot: One gene -> one entry (9 isoforms)

RefSeq: One mRNA sequence -> one entry



Results

1 out of 1 identifier from UniProtKB AC/ID was successfully mapped to 15 RefSeq Protein IDs.

From	To
P04637	NP_000537.3
P04637	NP_001119584.1
P04637	NP_001119585.1
P04637	NP_001119586.1
P04637	NP_001119587.1
P04637	NP_001119588.1
P04637	NP_001119589.1
P04637	NP_001119590.1
P04637	NP_001263624.1
P04637	NP_001263625.1
P04637	NP_001263626.1
P04637	NP_001263627.1
P04637	NP_001263628.1
P04637	NP_001263689.1
P04637	NP_001263690.1

ID/AC mapping: see later

Query: organism:"Homo sapiens (Human) [9606]"

Reviewed (20,171)
Swiss-Prot

Number of canonical and isoform protein sequences: 42,147 (download data in FASTA format)

Unreviewed (136,539)
TrEMBL

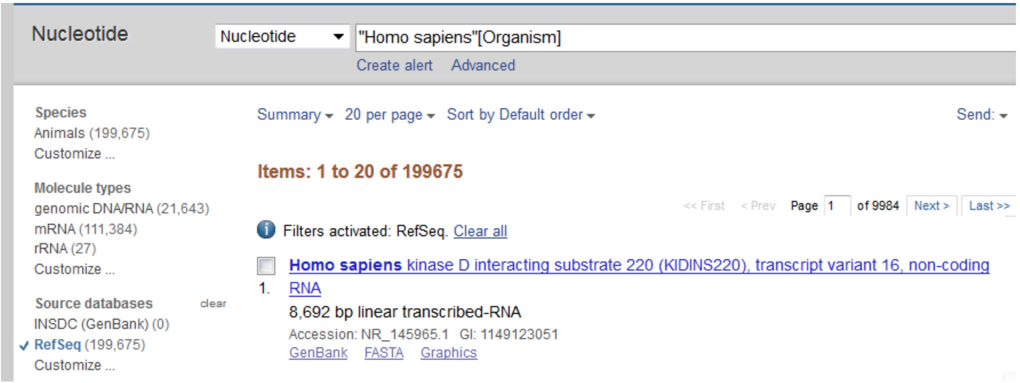
Popular organisms

Human (156,710) ✖

Proteomes

UP000005640 (70,952)

RefSeq



"Homo sapiens"[Organism]

Species
Animals (199,675)
Customize ...

Molecule types
genomic DNA/RNA (21,643)
mRNA (111,384)
rRNA (27)
Customize ...

Source databases clear
INSDC (GenBank) (0)
✓ RefSeq (199,675)
Customize ...

Items: 1 to 20 of 199675

Page 1 of 9984

Filters activated: RefSeq. Clear all

□ 1. Homo sapiens kinase D interacting substrate 220 (KIDINS220), transcript variant 16, non-coding RNA
8,692 bp linear transcribed-RNA
Accession: NR_145965.1 GI: 1149123051

UniProt

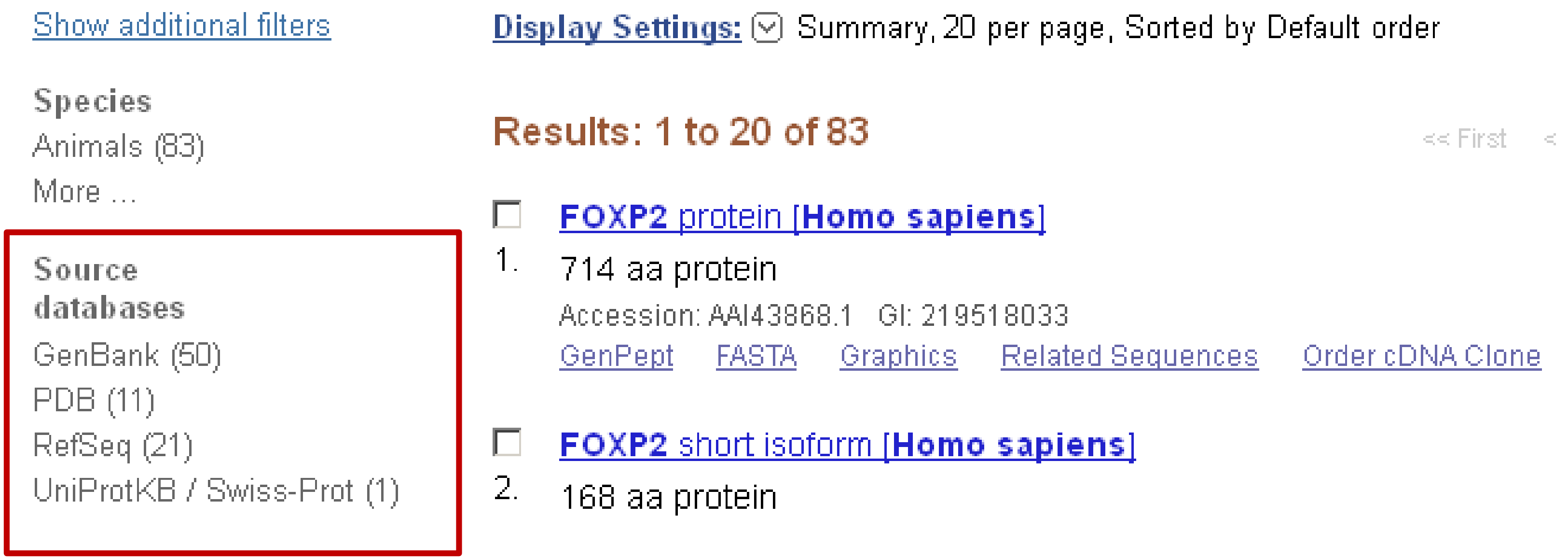
SIB Swiss Institute of Bioinformatics

NCBI nr

query & BLAST

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein

Look for human FOXP2 @ NCBI nr



Show additional filters

Species

Animals (83)

More ...

Source databases

GenBank (50)

PDB (11)

RefSeq (21)

UniProtKB / Swiss-Prot (1)

Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 83

1. □ **FOXP2** protein [**Homo sapiens**]

   714 aa protein

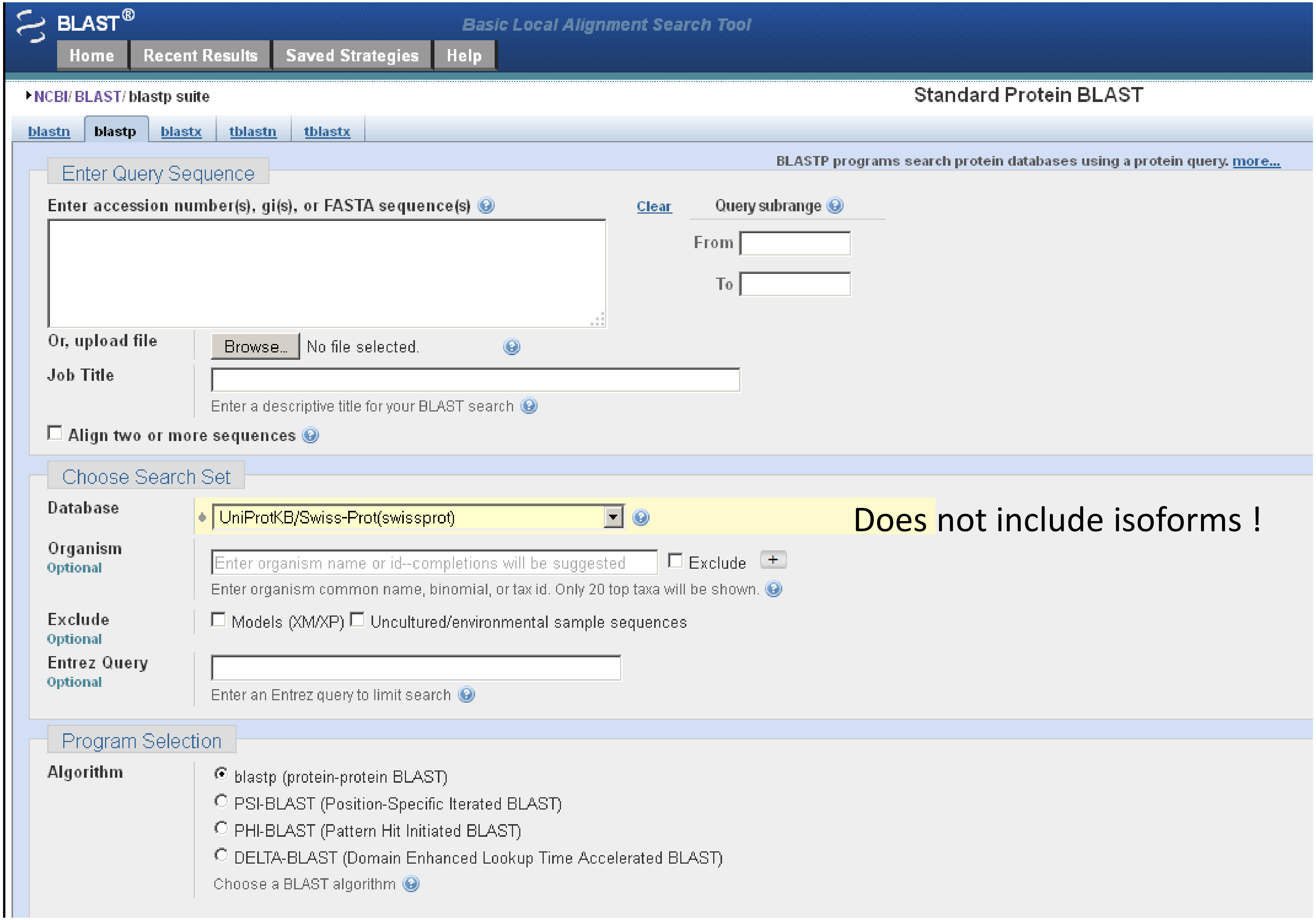
   Accession: AAI43868.1 GI: 219518033

   GenPept FASTA Graphics Related Sequences Order cDNA Clone

2. □ **FOXP2** short isoform [**Homo sapiens**]

   168 aa protein

BLAST @ NCBI



NCBI/ BLAST/ blastp suite

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Does not include isoforms !

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

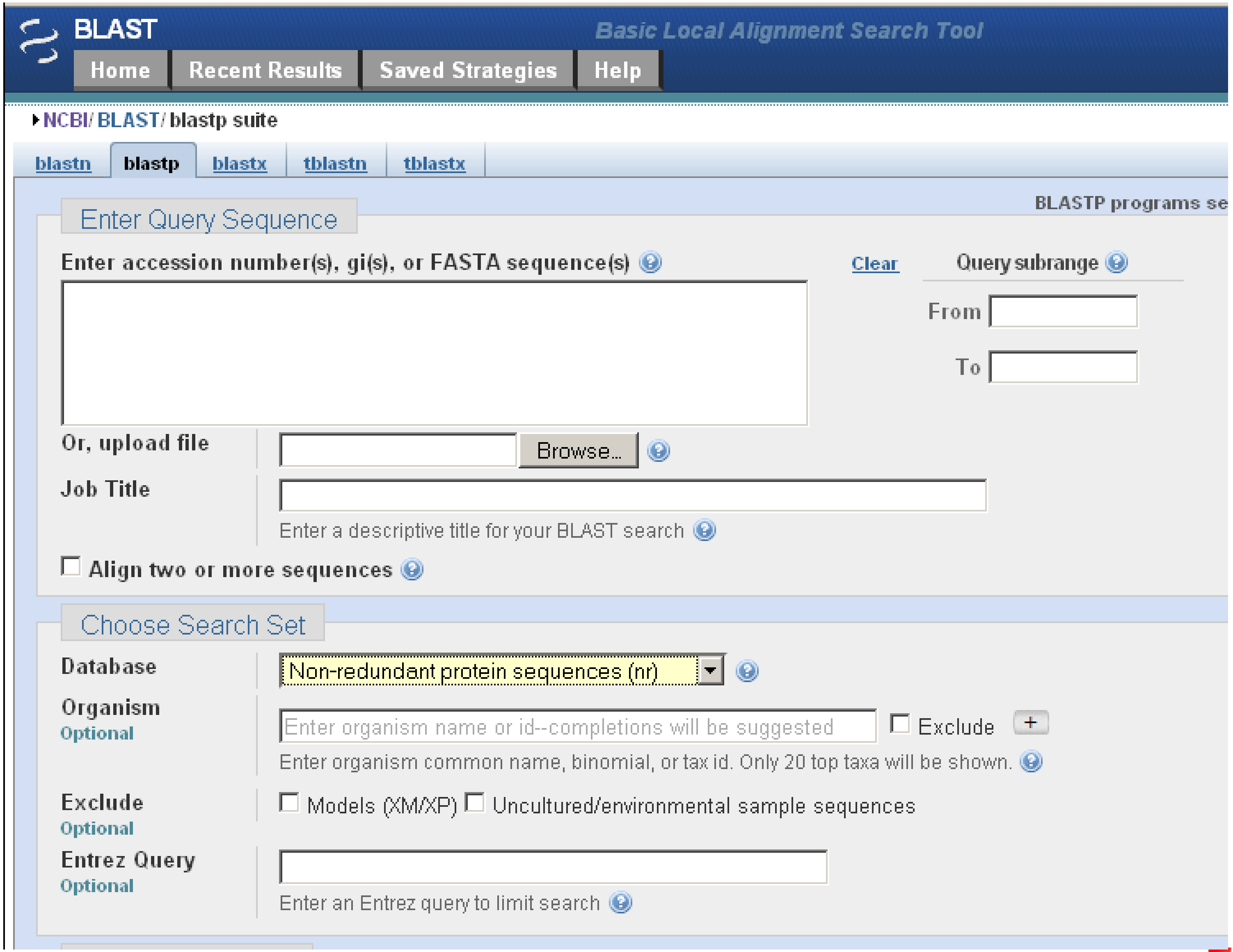
Program Selection

Algorithm

● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST @ NCBI



BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs se

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

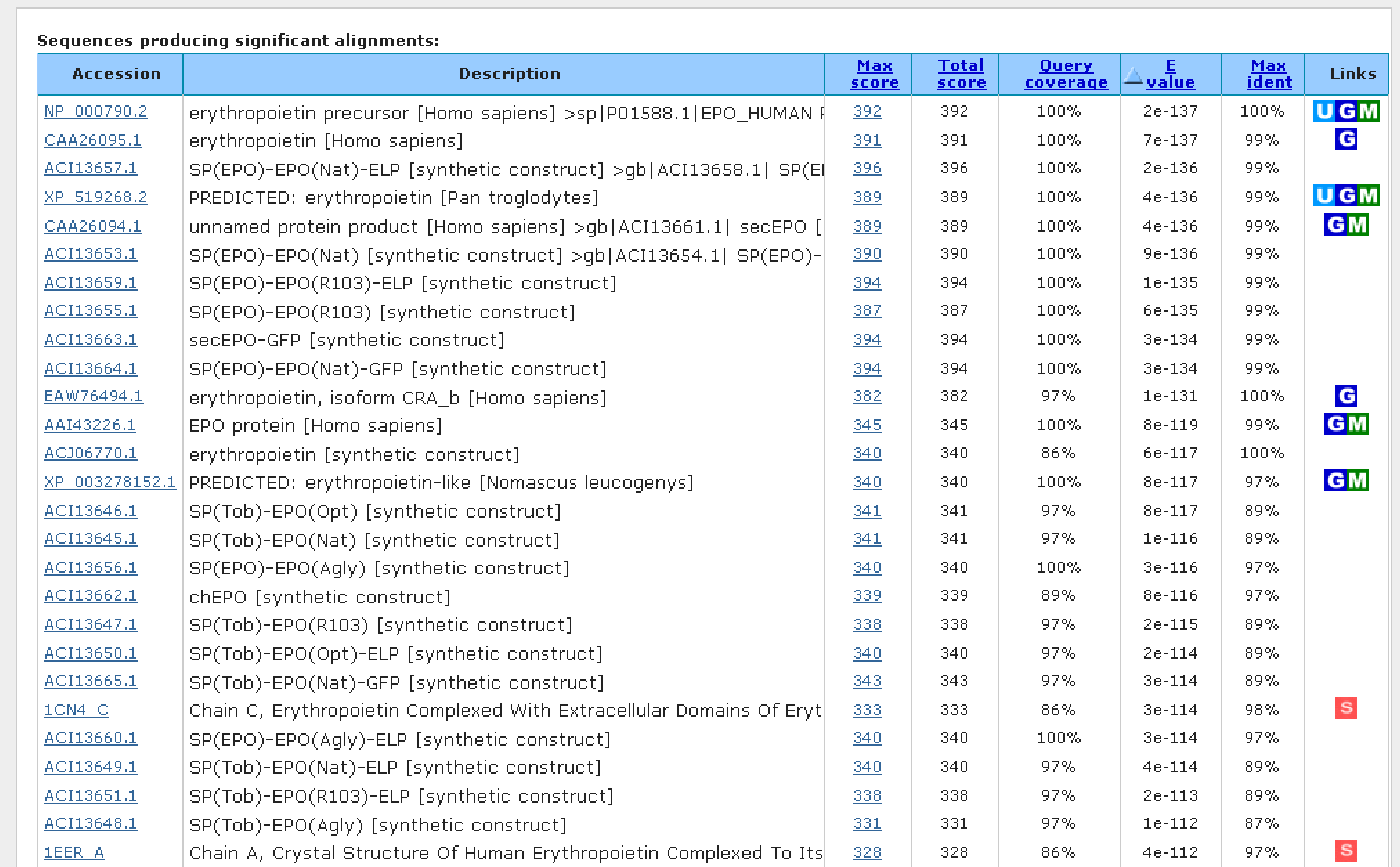
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

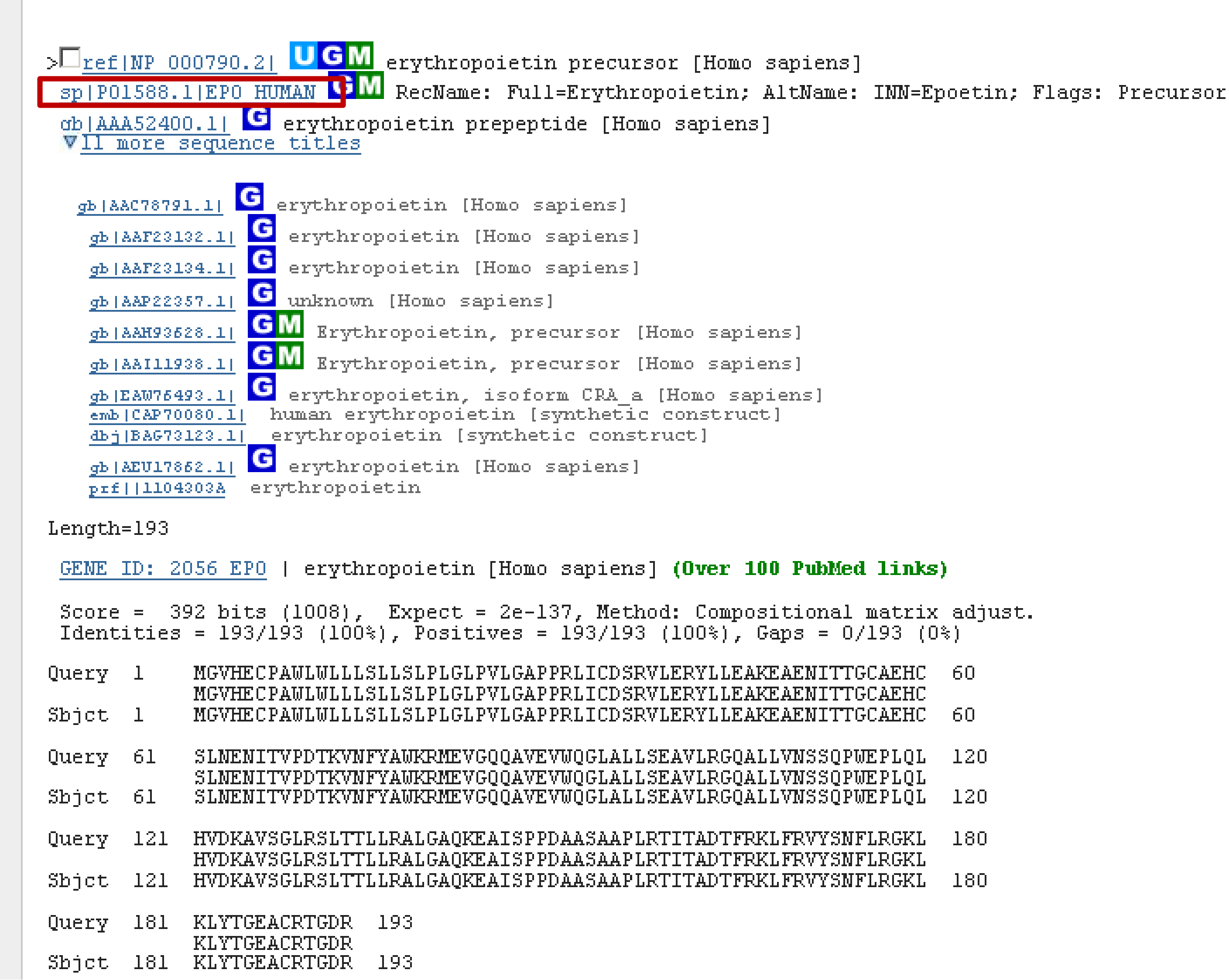


Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_000790.2	erythropoietin precursor [Homo sapiens] >sp|P01588.1|EPO_HUMAN F	392	392	100%	2e-137	100%	UGM
CAA26095.1	erythropoietin [Homo sapiens]	391	391	100%	7e-137	99%	G
ACI13657.1	SP(EPO)-EPO(Nat)-ELP [synthetic construct] >gb|ACI13658.1| SP(EI	396	396	100%	2e-136	99%	
XP_519268.2	PREDICTED: erythropoietin [Pan troglodytes]	389	389	100%	4e-136	99%	UGM
CAA26094.1	unnamed protein product [Homo sapiens] >gb|ACI13661.1| secEPO [	389	389	100%	4e-136	99%	GM
ACI13653.1	SP(EPO)-EPO(Nat) [synthetic construct] >gb|ACI13654.1| SP(EPO)-	390	390	100%	9e-136	99%	
ACI13659.1	SP(EPO)-EPO(R103)-ELP [synthetic construct]	394	394	100%	1e-135	99%	
ACI13655.1	SP(EPO)-EPO(R103) [synthetic construct]	387	387	100%	6e-135	99%	
ACI13663.1	secEPO-GFP [synthetic construct]	394	394	100%	3e-134	99%	
ACI13664.1	SP(EPO)-EPO(Nat)-GFP [synthetic construct]	394	394	100%	3e-134	99%	
EAW76494.1	erythropoietin, isoform CRA_b [Homo sapiens]	382	382	97%	1e-131	100%	G
AAI43226.1	EPO protein [Homo sapiens]	345	345	100%	8e-119	99%	GM
ACJ06770.1	erythropoietin [synthetic construct]	340	340	86%	6e-117	100%	
XP_003278152.1	PREDICTED: erythropoietin-like [Nomascus leucogenys]	340	340	100%	8e-117	97%	GM
ACI13646.1	SP(Tob)-EPO(Opt) [synthetic construct]	341	341	97%	8e-117	89%	
ACI13645.1	SP(Tob)-EPO(Nat) [synthetic construct]	341	341	97%	1e-116	89%	
ACI13656.1	SP(EPO)-EPO(Agly) [synthetic construct]	340	340	100%	3e-116	97%	
ACI13662.1	chEPO [synthetic construct]	339	339	89%	8e-116	97%	
ACI13647.1	SP(Tob)-EPO(R103) [synthetic construct]	338	338	97%	2e-115	89%	
ACI13650.1	SP(Tob)-EPO(Opt)-ELP [synthetic construct]	340	340	97%	2e-114	89%	
ACI13665.1	SP(Tob)-EPO(Nat)-GFP [synthetic construct]	343	343	97%	3e-114	89%	
1CN4_C	Chain C, Erythropoietin Complexed With Extracellular Domains Of Eryt	333	333	86%	3e-114	98%	S
ACI13660.1	SP(EPO)-EPO(Agly)-ELP [synthetic construct]	340	340	100%	3e-114	97%	
ACI13649.1	SP(Tob)-EPO(Nat)-ELP [synthetic construct]	340	340	97%	4e-114	89%	
ACI13651.1	SP(Tob)-EPO(R103)-ELP [synthetic construct]	338	338	97%	2e-113	89%	
ACI13648.1	SP(Tob)-EPO(Agly) [synthetic construct]	331	331	97%	1e-112	87%	
1EER_A	Chain A, Crystal Structure Of Human Erythropoietin Complexed To Its	328	328	86%	4e-112	97%	S

UniProtKB/Swiss-Prot alternative isoform sequences are not included !

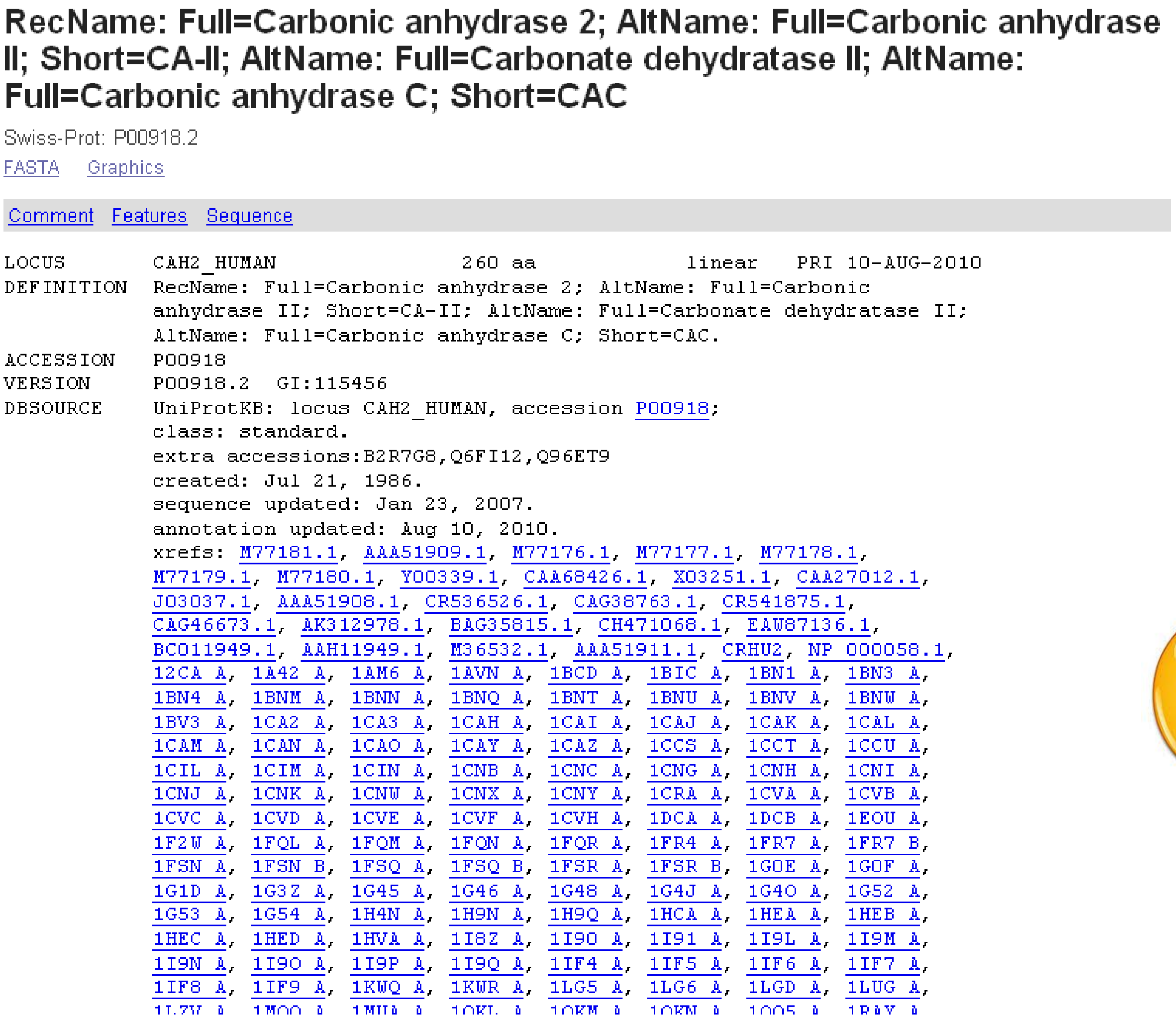
NCBI nr ‘cluster’ of identical sequences



```
>□ref|NP_000790.2| UGM erythropoietin precursor [Homo sapiens]
 sp|P01588.1|EPO_HUMAN UGM RecName: Full=Erythropoietin; AltName: INN=Epoetin; Flags: Precursor
 gb|AAA52400.1| G erythropoietin prepeptide [Homo sapiens]
 ▼11 more sequence titles

   gb|AAC78791.1| G erythropoietin [Homo sapiens]
    gb|AAF23132.1| G erythropoietin [Homo sapiens]
    gb|AAF23134.1| G erythropoietin [Homo sapiens]
    gb|AAP22357.1| G unknown [Homo sapiens]
    gb|AAH93628.1| GM Erythropoietin, precursor [Homo sapiens]
    gb|AAI11938.1| GM Erythropoietin, precursor [Homo sapiens]
    gb|EAW76493.1| G erythropoietin, isoform CRA_a [Homo sapiens]
    emb|CAP70080.1| human erythropoietin [synthetic construct]
    dbj|BAG73123.1| erythropoietin [synthetic construct]
    gb|AEU17862.1| G erythropoietin [Homo sapiens]
    prf||1104303A  erythropoietin

Length=193

 GENE ID: 2056 EPO | erythropoietin [Homo sapiens] (Over 100 PubMed links)

 Score =  392 bits (1008),  Expect = 2e-137, Method: Compositional matrix adjust.
 Identities = 193/193 (100%), Positives = 193/193 (100%), Gaps = 0/193 (0%)

Query  1    MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC  60
            MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC
Sbjct  1    MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC  60

Query  61   SLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQL  120
            SLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQL
Sbjct  61   SLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQL  120

Query  121  HVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRGKL  180
            HVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRGKL
Sbjct  121  HVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRGKL  180

Query  181  KLYTGEACRTGDR  193
            KLYTGEACRTGDR
Sbjct  181  KLYTGEACRTGDR  193
```

UniProtKB entries at NCBI…

A UniProtKB/Swiss-Prot entry with the NCBI look



RecName: Full=Carbonic anhydrase 2; AltName: Full=Carbonic anhydrase II; Short=CA-II; AltName: Full=Carbonate dehydratase II; AltName: Full=Carbonic anhydrase C; Short=CAC

Swiss-Prot: P00918.2

FASTA   Graphics

Comment   Features   Sequence

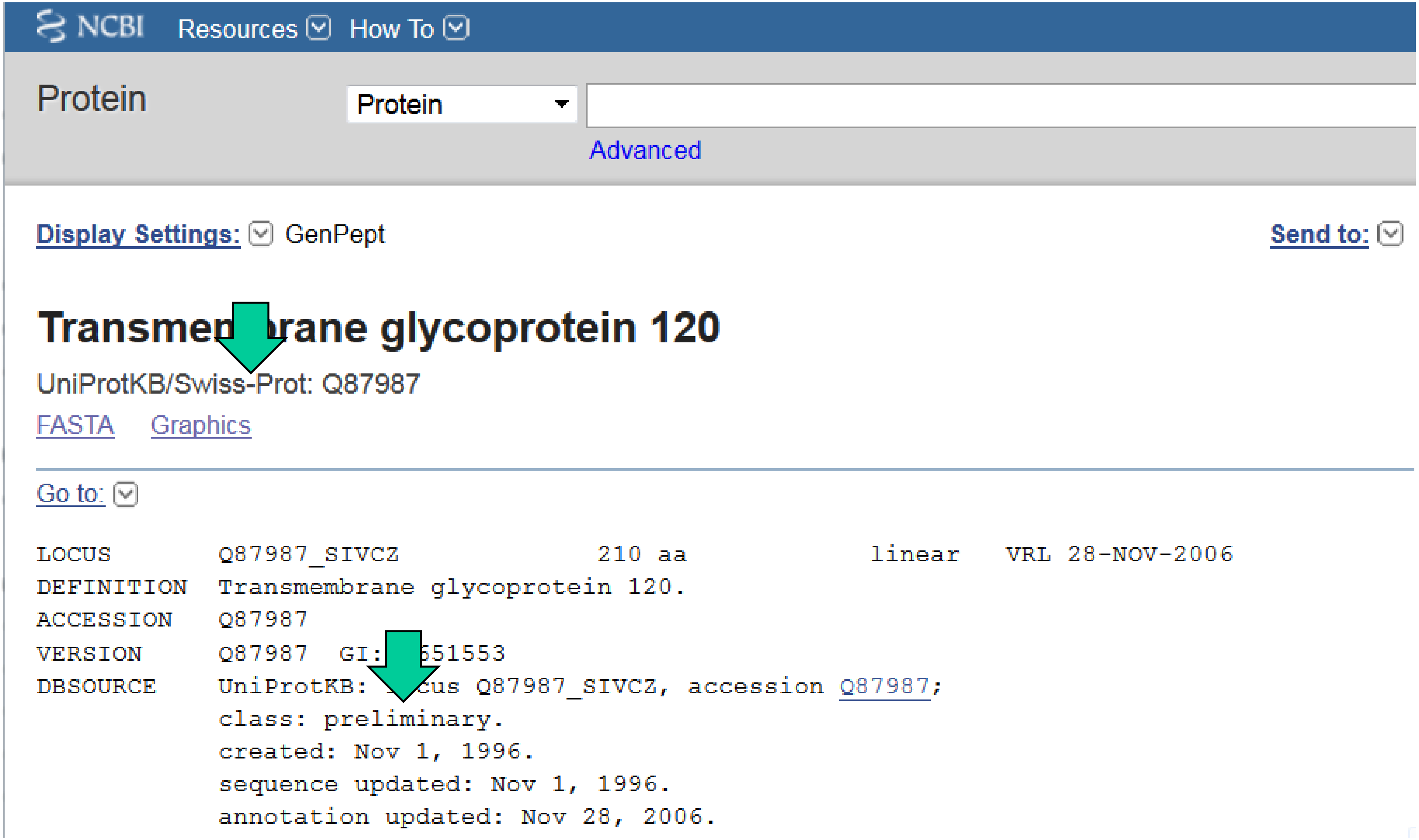
```
LOCUS       CAH2_HUMAN               260 aa            linear   PRI 10-AUG-2010
DEFINITION  RecName: Full=Carbonic anhydrase 2; AltName: Full=Carbonic
            anhydrase II; Short=CA-II; AltName: Full=Carbonate dehydratase II;
            AltName: Full=Carbonic anhydrase C; Short=CAC.
ACCESSION   P00918
VERSION     P00918.2  GI:115456
DBSOURCE    UniProtKB: locus CAH2_HUMAN, accession P00918;
            class: standard.
            extra accessions:B2R7G8,Q6FI12,Q96ET9
            created: Jul 21, 1986.
            sequence updated: Jan 23, 2007.
            annotation updated: Aug 10, 2010.
            xrefs: M77181.1, AAA51909.1, M77176.1, M77177.1, M77178.1,
            M77179.1, M77180.1, Y00339.1, CAA68426.1, X03251.1, CAA27012.1,
            J03037.1, AAA51908.1, CR536526.1, CAG38763.1, CR541875.1,
            CAG46673.1, AK312978.1, BAG35815.1, CH471068.1, EAW87136.1,
            BC011949.1, AAH11949.1, M36532.1, AAA51911.1, CRHU2, NP 000058.1,
            12CA A, 1A42 A, 1AM6 A, 1AVN A, 1BCD A, 1BIC A, 1BN1 A, 1BN3 A,
            1BN4 A, 1BNM A, 1BNN A, 1BNQ A, 1BNT A, 1BNU A, 1BNV A, 1BNW A,
            1BV3 A, 1CA2 A, 1CA3 A, 1CAH A, 1CAI A, 1CAJ A, 1CAK A, 1CAL A,
            1CAM A, 1CAN A, 1CAO A, 1CAY A, 1CAZ A, 1CCS A, 1CCT A, 1CCU A,
            1CIL A, 1CIM A, 1CIN A, 1CNB A, 1CNC A, 1CNG A, 1CNH A, 1CNI A,
            1CNJ A, 1CNK A, 1CNW A, 1CNX A, 1CNY A, 1CRA A, 1CVA A, 1CVB A,
            1CVC A, 1CVD A, 1CVE A, 1CVF A, 1CVH A, 1DCA A, 1DCB A, 1EOU A,
            1F2W A, 1FQL A, 1FQM A, 1FQN A, 1FQR A, 1FR4 A, 1FR7 A, 1FR7 B,
            1FSN A, 1FSN B, 1FSQ A, 1FSQ B, 1FSR A, 1FSR B, 1G0E A, 1G0F A,
            1G1D A, 1G3Z A, 1G45 A, 1G46 A, 1G48 A, 1G4J A, 1G4O A, 1G52 A,
            1G53 A, 1G54 A, 1H4N A, 1H9N A, 1H9Q A, 1HCA A, 1HEA A, 1HEB A,
            1HEC A, 1HED A, 1HVA A, 1I8Z A, 1I90 A, 1I91 A, 1I9L A, 1I9M A,
            1I9N A, 1I9O A, 1I9P A, 1I9Q A, 1IF4 A, 1IF5 A, 1IF6 A, 1IF7 A,
            1IF8 A, 1IF9 A, 1KWQ A, 1KWR A, 1LG5 A, 1LG6 A, 1LGD A, 1LUG A,
            1LZV A  1MOO A  1MUA A  1OKL A  1OKM A  1OKN A  1OQ5 A  1RAY A
```

Important remarks concerning the datasets

Different servers…

UniProtKB/TrEMBL entries are not available at NCBI
The same protein sequence might be present,
but not with the UniProtKB/TrEMBL AC (**with some exceptions**)
(not the case for UniProtKB/Swiss-Prot entries)

UniProtKB/TrEMBL entries are not available at NCBInr with some exceptions…



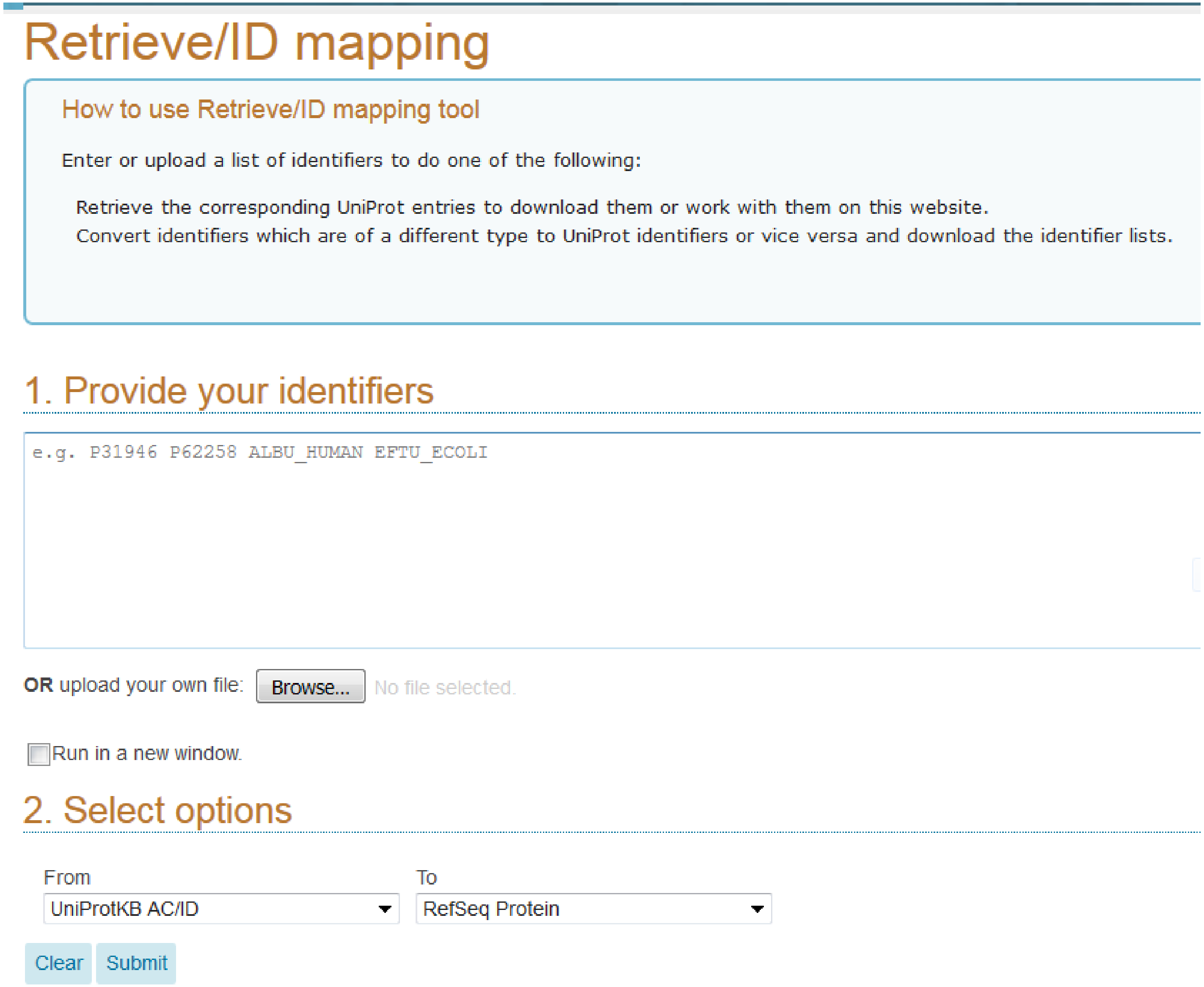
Transmembrane glycoprotein 120

UniProtKB/Swiss-Prot: Q87987

```
LOCUS       Q87987_SIVCZ              210 aa            linear   VRL 28-NOV-2006
DEFINITION  Transmembrane glycoprotein 120.
ACCESSION   Q87987
VERSION     Q87987  GI:  651553
DBSOURCE    UniProtKB: locus Q87987_SIVCZ, accession Q87987;
            class: preliminary.
            created: Nov 1, 1996.
            sequence updated: Nov 1, 1996.
            annotation updated: Nov 28, 2006.
```

ID/AC mapping

http://www.uniprot.org/uploadlists/



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

e.g. P31946 P62258 ALBU_HUMAN EFTU_ECOLI

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

□ Run in a new window.

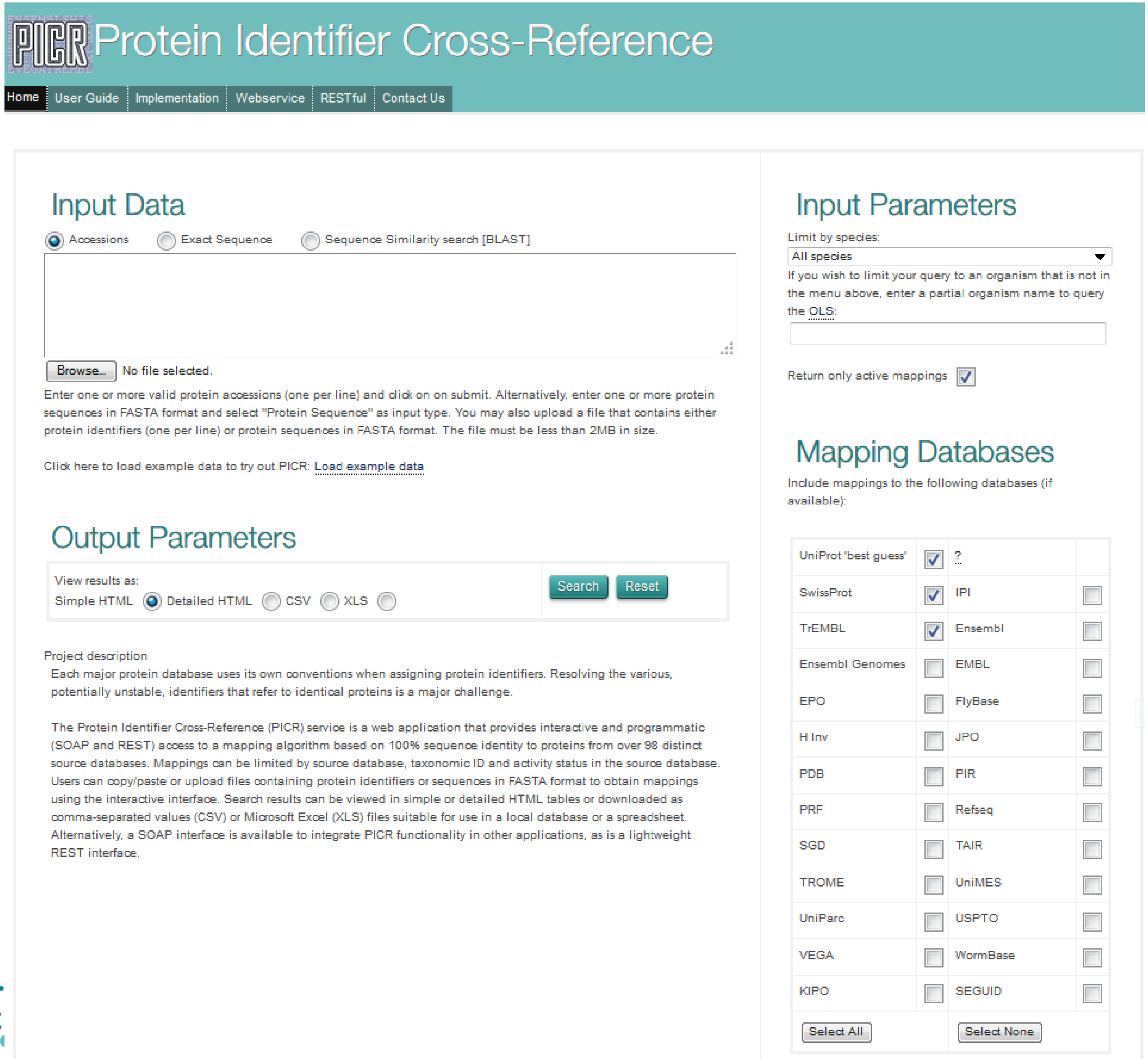
2. Select options

From: UniProtKB AC/ID

To: RefSeq Protein

Swiss Institute of Bioinformatics

http://www.ebi.ac.uk/Tools/picr/



PICR Protein Identifier Cross-Reference

Input Data

● Accessions ○ Exact Sequence ○ Sequence Similarity search [BLAST]

Browse... No file selected.

Enter one or more valid protein accessions (one per line) and click on submit. Alternatively, enter one or more protein sequences in FASTA format and select "Protein Sequence" as input type. You may also upload a file that contains either protein identifiers (one per line) or protein sequences in FASTA format. The file must be less than 2MB in size.

Click here to load example data to try out PICR: Load example data

Output Parameters

View results as:
Simple HTML ● Detailed HTML ○ CSV ○ XLS ○

Project description

Each major protein database uses its own conventions when assigning protein identifiers. Resolving the various, potentially unstable, identifiers that refer to identical proteins is a major challenge.

The Protein Identifier Cross-Reference (PICR) service is a web application that provides interactive and programmatic (SOAP and REST) access to a mapping algorithm based on 100% sequence identity to proteins from over 98 distinct source databases. Mappings can be limited by source database, taxonomic ID and activity status in the source database. Users can copy/paste or upload files containing protein identifiers or sequences in FASTA format to obtain mappings using the interactive interface. Search results can be viewed in simple or detailed HTML tables or downloaded as comma-separated values (CSV) or Microsoft Excel (XLS) files suitable for use in a local database or a spreadsheet. Alternatively, a SOAP interface is available to integrate PICR functionality in other applications, as is a lightweight REST interface.

Input Parameters

Limit by species:
All species

If you wish to limit your query to an organism that is not in the menu above, enter a partial organism name to query the OLS:

Return only active mappings ☑

Mapping Databases

Include mappings to the following databases (if available):

UniProt 'best guess'	☑	?	
SwissProt	☑	IPI	☐
TrEMBL	☑	Ensembl	☐
Ensembl Genomes	☐	EMBL	☐
EPO	☐	FlyBase	☐
H Inv	☐	JPO	☐
PDB	☐	PIR	☐
PRF	☐	Refseq	☐
SGD	☐	TAIR	☐
TROME	☐	UniMES	☐
UniParc	☐	USPTO	☐
VEGA	☐	WormBase	☐
KIPO	☐	SEGUID	☐
Select All		Select None	

UniProt

SIB Swiss Institute of Bioinformatics

Understanding protein function is critical to research in many areas of science such as biology, medicine and biotechnology.

Keeping up with all of this information is a daunting task for most researchers.
UniProt helps with this in the following ways:

- it provides an **up-to-date**, comprehensive body of protein information at a single site;
- it aids scientific discovery by collecting, **interpreting and organising** this information so that it is easy to access and use;
- it saves researchers countless hours of work in monitoring and **collecting** this information themselves;
- it provides tools to help with **protein sequence analysis**;
- it **provides links** to related information in more than 150 other biological databases to help you access additional information in more specialised collections.

- https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional/why-do-we-need-uniprot

Thank you !

Thanks to Emmanuel Boutet and Marc Feuermann
for some of the slides !

Thanks to Diana Marek & Geoff Fucile
for the organisation of this course

Nucleic acid sequence databases

- **INSDC, RefSeq, Ensembl**

Protein sequence databases

UniProtKB

- **UniProtKB/Swiss-Prot**
  - **Protein sequences**
  - **Biological knowledge**
- **UniProtKB/TrEMBL**
  - **Protein sequences**
  - **Biological knowledge**
- **GO annotation**
- **UniProt Proteomes**

NCBI protein (RefSeq)

Practicals