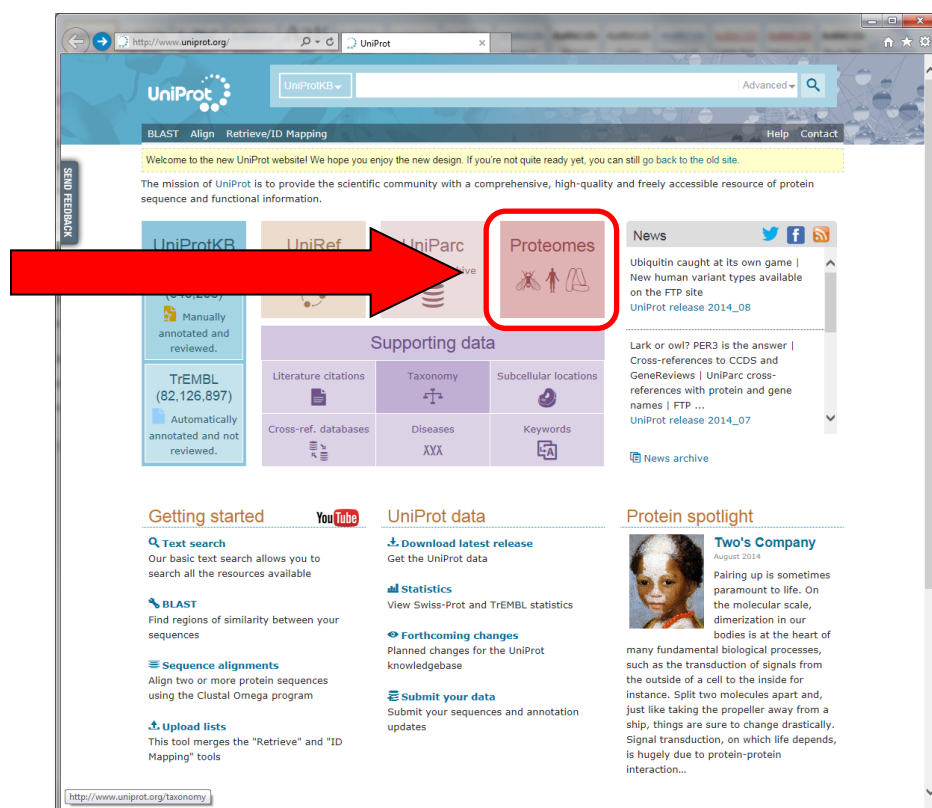


# Database Generation

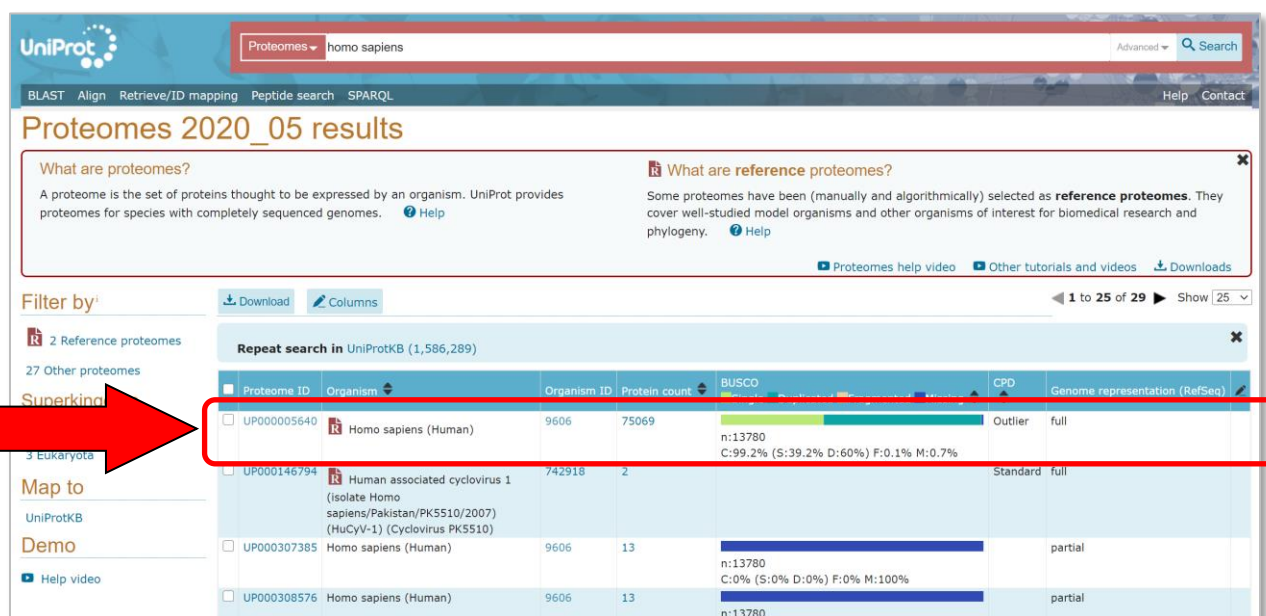
In order to identify peptides and proteins, we are going to compare the mass spectra to *in silico* theoretic spectra deduced from a protein database. *Are there other database types that could be used to identify the spectra? Would it even be possible to identify the spectra without a database at all?* [1.1a]

The choice of the database is crucial for the identification procedure. Indeed, shotgun proteomics workflows will only retrieve proteins contained in the database: the database should contain all possible sequences. Yet, if the database is too large, the search engine will have more room for mistakes and will introduce false positive identifications. The UniProt<sup>1</sup> database is a repository of choice for proteomics as it allies quality and quantity of protein sequences.

In order to optimize the database size, we will select only the species needed. The spectra in our example were obtained from a human sample. Go to the UniProt website ([www.uniprot.org](http://www.uniprot.org)) and select “Proteomes” in the middle of the screen.



Next fill in **homo sapiens** in the “Proteomes” search field at the top and hit enter. UniProt retrieves 29 proteomes (as of June 2021), including two reference proteomes. Select the “Homo sapiens (Human)” result marked as a reference proteome.



UniProt

Proteomes

BLAST Align Retrieve/ID mapping Peptide search SPARQL Help Contact

## Proteomes 2020\_05 results

What are proteomes?  
A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes. [Help](#)

What are reference proteomes?  
Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. [Help](#)

Proteomes help video Other tutorials and videos Downloads

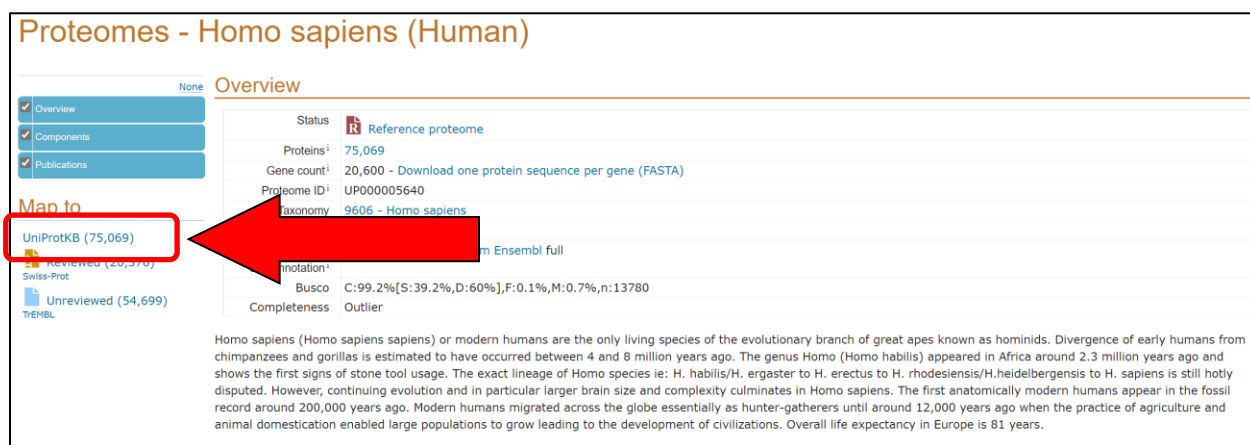
Filter by:   1 to 25 of 29 Show 25

2 Reference proteomes  
27 Other proteomes

Repeat search in UniProtKB (1,586,289)

Proteome ID	Organism	Organism ID	Protein count	BUSCO	CPD	Genome representation (RefSeq)
<input type="checkbox"/> UP000005640	Homo sapiens (Human)	9606	75069	n:13780 C:99.2% (S:39.2% D:60%) F:0.1% M:0.7%	Outlier	full
<input type="checkbox"/> UP000146794	Human associated cyclovirus 1 (isolate Homo sapiens/Pakistan/PK5510/2007) (HuCyV-1) (Cyclovirus PK5510)	742918	2		Standard	full
<input type="checkbox"/> UP000307385	Homo sapiens (Human)	9606	13	n:13780 C:0% (S:0% D:0%) F:0% M:100%		partial
<input type="checkbox"/> UP000308576	Homo sapiens (Human)	9606	13	n:13780		partial

Then select UniProtKB option to the left:



## Proteomes - Homo sapiens (Human)

None Overview

☒ Overview  
☒ Components  
☒ Publications

Map to

Status: Reference proteome

Proteins: 75,069

Gene count: 20,600 - Download one protein sequence per gene (FASTA)

Proteome ID: UP000005640

Taxonomy: 9606 - Homo sapiens

Annotation: BUSCO: C:99.2% (S:39.2%, D:60%) F:0.1%, M:0.7%, n:13780  
Completeness: Outlier

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

This will show you all the proteins in the human reference proteome in UniProt:

UniProtKB 2020\_05 results

Filter by: Reviewed (20,370) Swiss-Prot, Unreviewed (54,699) TrEMBL, Popular organisms: Human (75,069), Proteomes: UP000005640 (75,069)\*

View by: Results table, Taxonomy, Keywords, Gene Ontology, Enzyme class, Pathway

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q8N7X0	ADGB_HUMAN	Androglobin	ADGB C6orf103, CAPN7L	Homo sapiens (Human)	1,667
<input type="checkbox"/> Q5T1N1	AKND1_HUMAN	Protein AKNAD1	AKNAD1 C1orf62	Homo sapiens (Human)	836
<input type="checkbox"/> Q92667	AKAP1_HUMAN	A-kinase anchor protein 1, mitochon...	AKAP1 AKAP149, PRKA1	Homo sapiens (Human)	903
<input type="checkbox"/> Q5VUY0	ADCL3_HUMAN	Arylacetamide deacetylase-like 3	AADACL3	Homo sapiens (Human)	407
<input type="checkbox"/> P62736	ACTA_HUMAN	Actin, aortic smooth muscle	ACTA2 ACTSA, ACTVS, GIG46	Homo sapiens (Human)	377
<input type="checkbox"/> Q9H553	ALG2_HUMAN	Alpha-1,3/1,6-mannosyltransferase A...	ALG2 UNQ666/PRO1298	Homo sapiens (Human)	416
<input type="checkbox"/> P0C7M7	ACSM4_HUMAN	Acyl-coenzyme A synthetase ACSM4, m...	ACSM4	Homo sapiens (Human)	580
<input type="checkbox"/> P49703	ARL4D_HUMAN	ADP-ribosylation factor-like protei...	ARL4D ARF4L	Homo sapiens (Human)	201
<input type="checkbox"/> Q5TGY3	AHDC1_HUMAN	AT-hook DNA-binding motif-containin...	AHDC1	Homo sapiens (Human)	1,603
<input type="checkbox"/> Q75V66	ANO5_HUMAN	Anoctamin-5	ANO5 GDD1, TMEM16E	Homo sapiens (Human)	913
<input type="checkbox"/> Q96JD6	AKCL2_HUMAN	1,5-anhydro-D-fructose reductase	AKR1E2 AKR1CL2, AKRDC1	Homo sapiens (Human)	320

How many proteins can we find for the human proteome? How is the protein sequence list established? Is it exhaustive? What is the difference between a reviewed and an unreviewed entry? [1.1b]

Select the option "Reviewed (Swiss-Prot)" in the upper left corner to list only the reviewed proteins and then click the "Download" button above the table. Make sure that "Download all" is selected, the format is set to "FASTA (canonical)" and "Uncompressed" is selected. Click "Go" to start downloading the protein sequences, and save the file as `uniprot-human-reviewed-june-2021.fasta`.

What is a canonical sequence? And what is the difference between the options 'FASTA (canonical)' and 'FASTA (canonical & isoform)'? [1.1c]



**Tips:**

*Always document your database type and version in the file name!*

*Organize your databases in a meaningful way for you and your colleagues!*

We now have a FASTA file with all the reviewed human protein sequences. However, our sample may also contain proteins from other species than the one we are studying, most often as a result of sample contamination.

Considering sample contamination is especially important when searching non-human data, as minute amounts of human keratin, from hair or skin, often end up in the samples. If these are not filtered out as contaminants, the search engines may very well mistake them as evidence for proteins not actually in the sample<sup>2</sup>. A list of common contaminants can be found at the Global Proteome Machine<sup>3</sup> (GPM) website (<http://www.thegpm.org/crap>).

As we are here working with human data we do not have to add human keratin to our list of proteins (there may be keratin from other species though). *Why not?* [1.1d]

However, there is still one non-human protein that can be detected in our sample, and that is the enzyme we used to digest the proteins into peptides. To reduce the chances of peptides from the digestion enzyme being used as evidence for the proteins actually in the sample we will therefore add the protein sequence of Trypsin to our FASTA file.



Go back to the main [UniProt](https://www.uniprot.org) webpage and search for the Trypsin protein from pig: "P00761".

**P00761 - TRYP\_PIG**  
 Protein: Trypsin  
 Gene: N/A  
 Organism: *Sus scrofa* (Pig)  
 Status: Reviewed - Experimental evidence at protein level<sup>1</sup>

**Display** None | BLAST | Align | Format | Add to basket | History | Comment (0) | Feedback | Help video

**Function<sup>1</sup>**  
 Catalytic activity<sup>1</sup>  
 Preferential cleavage: Arg-|-Xaa, Lys-|-Xaa.  
 Cofactor<sup>1</sup>  
 Binds 1 calcium ion per subunit.  
 Sites

Feature key	Position (s)	Length	Description	Graphical view	Feature identifier	Actions
Active site <sup>1</sup>	48 - 48	1	Charge relay system			
Metal binding <sup>1</sup>	60 - 60	1	Calcium			
Metal binding <sup>1</sup>	62 - 62	1	Calcium; via carbonyl oxygen			
Metal binding <sup>1</sup>	65 - 65	1	Calcium; via carbonyl oxygen			
Metal binding <sup>1</sup>	70 - 70	1	Calcium			
Active site <sup>1</sup>	92 - 92	1	Charge relay system			
Site <sup>1</sup>	179 - 179	1	Required for specificity By similarity			
Active site <sup>1</sup>	185 - 185	1	Charge relay system			

**GO - Molecular function<sup>1</sup>**  
 • metal ion binding Source: UniProtKB-KW  
 • serine-type endopeptidase activity Source: InterPro

**GO - Biological process<sup>1</sup>**  
 • digestion Source: UniProtKB-KW  
 Complete GO annotation...

**Keywords - Molecular function<sup>1</sup>**

Click the "Sequence" option in the left menu and click the FASTA download option:

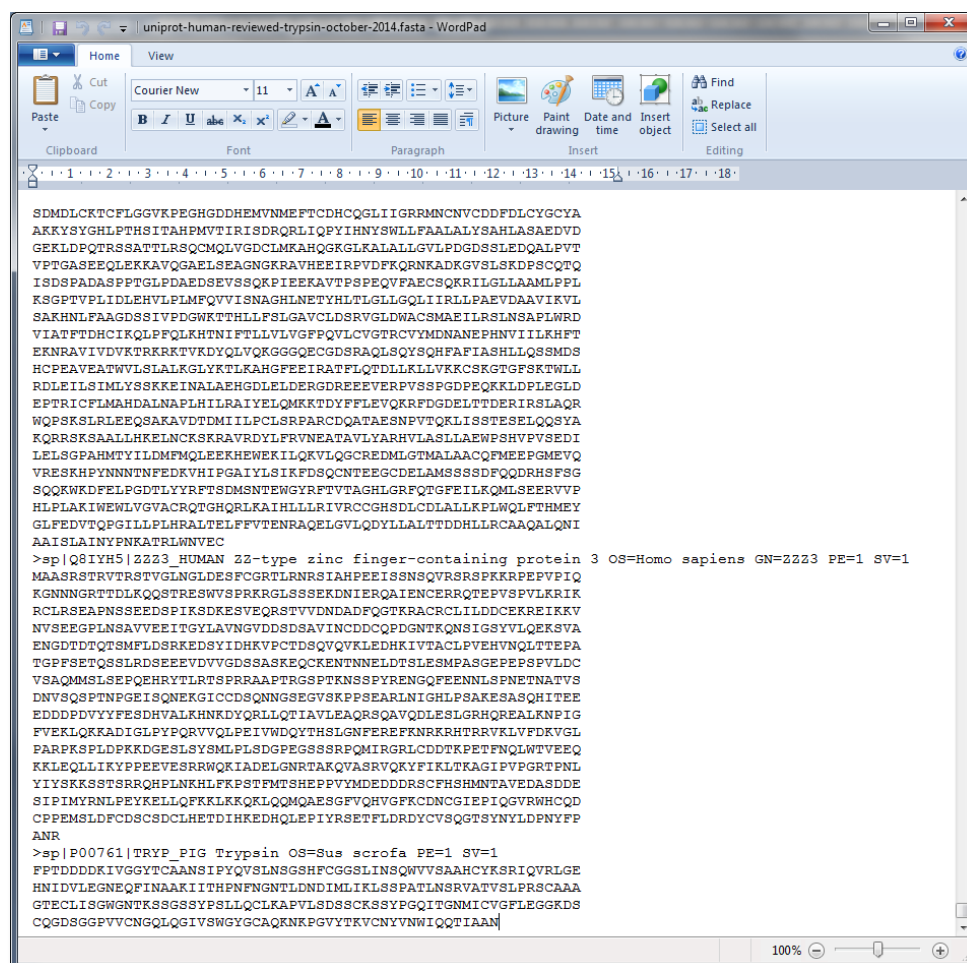
```
>sp|P00761|TRYP_PIG Trypsin OS=Sus scrofa PE=1 SV=1
FFIDDDKIVGGITCAANSIPYQVSLNSGSHFCGGSLNSGWVYSAHCYKSRIGVRLGE
HHIDVLEGGEGFIAAKIITPPIFPPTLNDIHLIKLSPATLNSRVATVSLPSCAAA
GTECLISGKNTKSSGSSVPSLLQCLKAPVLSDSSCKSSYFGQITGNMTCVFLGGKDS
CQSDSGGPVVCNGQLQGISVNGYGCAQKNKFGVYTKVCHYVNWIQQTIAAN
```

Next, double click the just created `uniprot-human-reviewed-june-2021.fasta` file to open it in a text editor such as WordPad (NotePad is not recommended as .fasta files can be large).



Double clicking the file may result in a dialog telling you that the file type/format is not recognized and ask you to select which program to open it in. This will only happen the first time you open a .fasta file.

Scroll to the bottom of the file, and paste in the Trypsin sequence. Make sure to not alter the formatting of the file or any of the sequence details.



```

SDMDLCKTCFLGGVKEPGHGDHDMVNMEFTCDHCQGLIIGRMNCNVCDPDFLCYGCYA
AKKYSYGHLPHTSITAHFMTIRISDRQLIQPYIHNYSWLLFAALALYSAHLASAEVD
GEKLDPQTRSSATTLRSQCMQLVGDCLMKAHQGGKGLKALALLGVLDGDSLEDQALPVT
VPTGASEEQLEKKAVQGAELSEAGNGKRAVHEEIRPVDKQRNKADKGVSLSKDPSCQTQ
ISDSPADASPPTGLPDAEDSEVSSQKPIEEKAVTPSPQVFAECSSQKRILGLLAAMLPL
KSGTVPVLIDLEHVLPLMFQVVISNAGHINETYHLLTGLLQQLIIRLLPAEVDAAIKVL
SAKHNLFAGDSSIVPDGKTKTHLLFSLGAVCLDSRVGLDWACSMAEILASINSAPLWRD
VIATFTDHCIKQLPFLKHTNIFTLLVLVGFPPQLVCGTRCVYMDNANEPHNVIILKHT
EKNRAVIVDVTKRRKTKVDYQLVQGGGQECGDSRAQLSQYSQHFAFIASHLLQSSMDS
HCPEAVEATWVLSLALKGLYKTLKAHGFEERATFLQTDLLKLLVKCKSGTGFSTKTLWL
RDLEILSLIMLYSSKKEINALAEHGDLELDERGDREEVERPVSSPGDPEQKKLDPLEGLD
EPTRICFLMAHDALNAPLHILRAIYELQMKKTDYFFLEVQKRFDGDELTTDERIRSLAQ
WQPSKSLRLEEQSAKAVDTDMIILPCLSRPARCDQATAESNPVTQKLISSESELQSSYA
KQRRSKSAALLHKELNCKSKRAVRDYLFRVNEATAVLYARHVLASLLAEWPSHVPSEDI
LELSPAHMTYILDMFMQLKEEKHEWEKILQKVLQGCREDMLGTMLAACQFMEEPMGEVQ
VRESKHPYNNNTNFEDKVIHPGAIIYLSIKFDSQCNTTEEGCDELAMSSSSDFQQDRHSFSG
SQQKKWDFELPGDGLTYRFTSDMSNTENGWYRFTVTAGHLGRFQTGFELKQMLSEERVFP
HLPLAKIWEHLVGVACRQTGHQRLKAHLLLRIVRCCGHSDDLCLALLKPLNQLFTHMEY
GLFEDVTQPGILLPLHRAITELFFVTENRAQELGLVLDYLLALTDDHLLRCAAQALQNI
AAISLAINYPNKATRLWNVEC
>sp|Q8IYH5|Z223_HUMAN Z2-type zinc finger-containing protein 3 OS=Homo sapiens GN=Z223 PE=1 SV=1
MAASRSTRVTRSTVGLNGLDESFCGRITLNRNSIAHPPEEISSNSQVRSRSPKKRPEPVPIQ
KGNNGRRTDLDKQSTRESWVSPRKRGLSSSEKDNIERQAIENCERRQTEPVSPVLKRIK
RCRLSEAPNSSEEDSPIKSDKESVEQRSTVVDNDADFQGTQRACRCLILDDCEKREIKKV
NVSESGPLNSAVVEEITGYLAVNGVDDSDSAVINCDCCQPDGNTKNSIGSYVLQKESVA
ENGDDTDTQSMFLDSRKEDSYIDHKVPCDTSQVQVLELHKIVTACLVEHVNLQLTPEPA
TGPFSSETQSSLRDSEEEVDVVGDSASKEQCKENTNNELDTSLEMPASGEPEPSPVLDC
VSAQMSLSEPEHRYTLRTSPRAAPTRGSPCKTNSPYRENGQFEENNLSPNETNATVS
DNVSQSPTNPGEISQNEKGICCDSCQNGSGVSKPPSEARLNIGHLPSAKESASQHITEE
EDDDPDVYVYFESDHVALKHNKDYQLRLQTIQVLEAQRQAVQDLESLSGRHQRREALKNPIG
FVEKQLKADIGLPPQRVVQLPEIVWDQYTHSLGNFEREFNRKRHTRRVKLVFDKVL
FARPKSFLDFKKGESLSYSMLPLSDGPEGSSSRPQMIRGLCDDTKPFETFNQLWTVEEQ
AKLEQLLIKYPPEEVESRRWQAIADLGNRTAKQVASRVQYFIKLTAKGIPVPGRTPNL
YIYSKSSSTRRRQHLNKLHFKPSTFMTSHEPPVYMDDDDRSCFHSMTAVEDASDDE
SIFIMYRNLPYKELLQFKKLLKQKLQMQAEGGFVQHVGFKCDNCGIEPIQGVWRHQCQD
CPPEMSLDFCDSCDCLHETDIHKEDHQLEPIYRSETFLDRDYCVSGTSSYNLYDNPYFP
ANR
>sp|P00761|TRYF_PIG Trypsin OS=Sus scrofa PE=1 SV=1
FPFDDDDKIVGGYTCAANSIPYQVSLNSGSHFCGGSLSNQWVSAACHYKSRIQVRLGE
HNIDVLGNEQFINAAKIITHFNFNGNTLDNDIMLIKLSPPATLNSRVATVSLPRSCAAA
GTCECLISGWGNTKSSGSSYPSSLQCLKAPVLSDSKSSYPGGITGNMICVGFLEGGKDS
CQGDGSGPVVVCNGQLQGIIVSGYGCAQKKNKPGVYTRVCNVYNNWIQQTIAAN
  
```

Save this new file as **uniprot-human-reviewed-trypsin-june-2021.fasta**.

You now have the desired FASTA file required to search the mass spectrometry example dataset.



This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 License.  
Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

# Advanced – Non-Standard Databases

For some studies, one has to create a non-standard database. This is facilitated by the relatively simple syntax of the FASTA format which can be edited in a normal text editor:

```
>header  
SEQUENCE
```

As illustrated here with the sequence of a human protein:

```
>sp|A6NCN2|K121P_HUMAN Keratin-81-like protein KRT121P OS=Homo sapiens GN=KRT121P PE=5 SV=4  
MEANSGRLASELNHVQEVLEGYKKKYEEVALRATAENEFVALKKDVCAYLRKSDLEAN  
VEALTQEIDFLRRLYEEEIRVLQSHISDTSVVVKMDNSRDLNMHCVITEIKAQYDDIATR  
SRAEASWYRSKCEEMKATVIRHGETLRRTKEEINELNRMIQRLTAEVENAKCQNSKLEA  
AVAQSEQQGAAALSDARCKLAELEGALQKAKQDMACLIREYQEVMSKLAWTLRSPPTGA  
CWRARSRGCVRALVL
```

It is however vital that the syntax used for the header is compatible with the search engines and the tools used to process the search results. For homemade databases, we recommend a generic format as detailed on our Database Help page (<http://compomics.github.io/projects/searchgui/wiki/DatabaseHelp.html>). There you will find information about how to set up your own custom databases.



---

# References

---

1. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119 (2004).
2. Ghesquiere, B., Helsens, K., Vandekerckhove, J. & Gevaert, K. A stringent approach to improve the quality of nitrotyrosine peptide identifications. *Proteomics* **11**, 1094-1098 (2011).
3. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **3**, 1234-1242 (2004).

