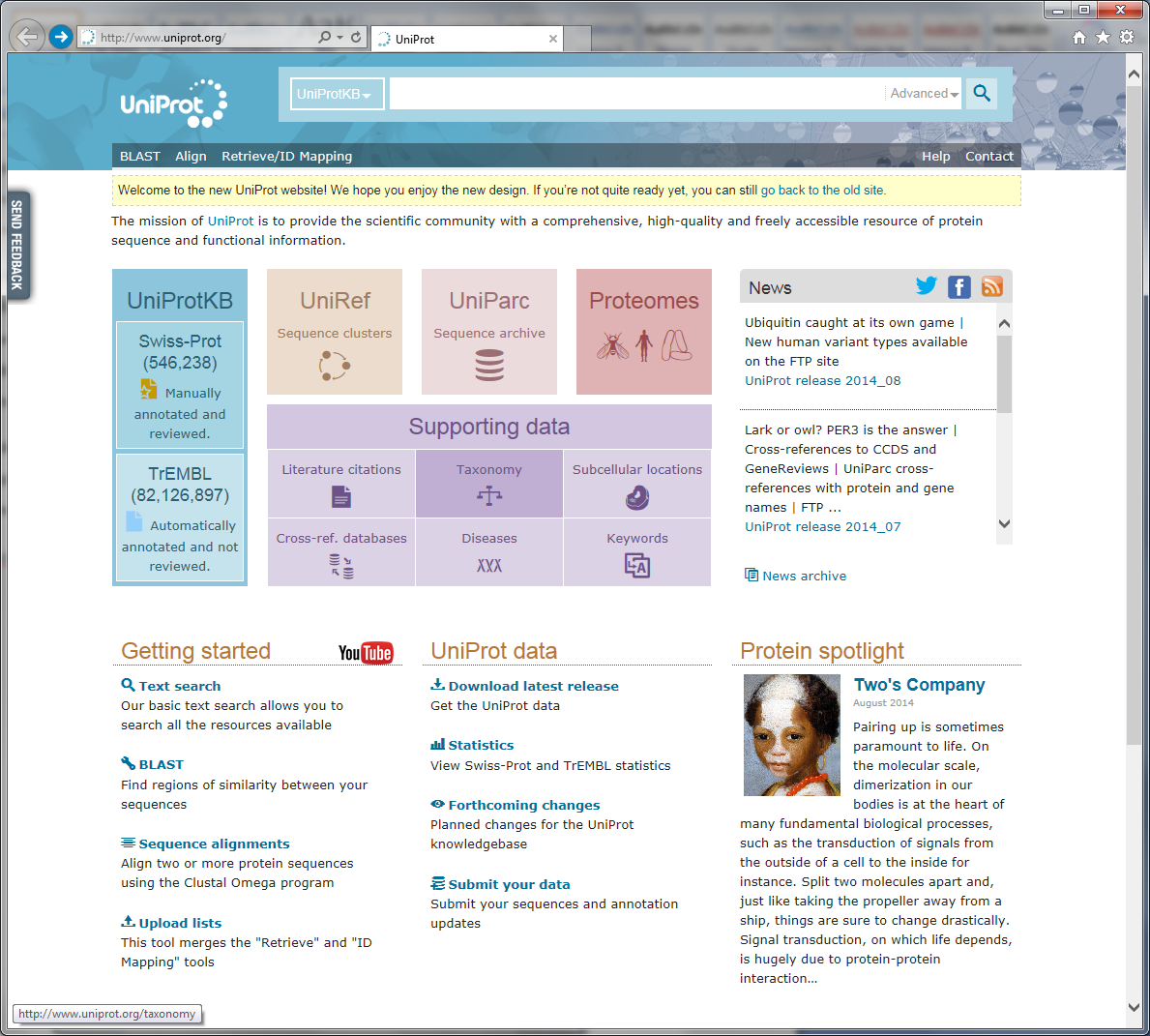
Database Generation

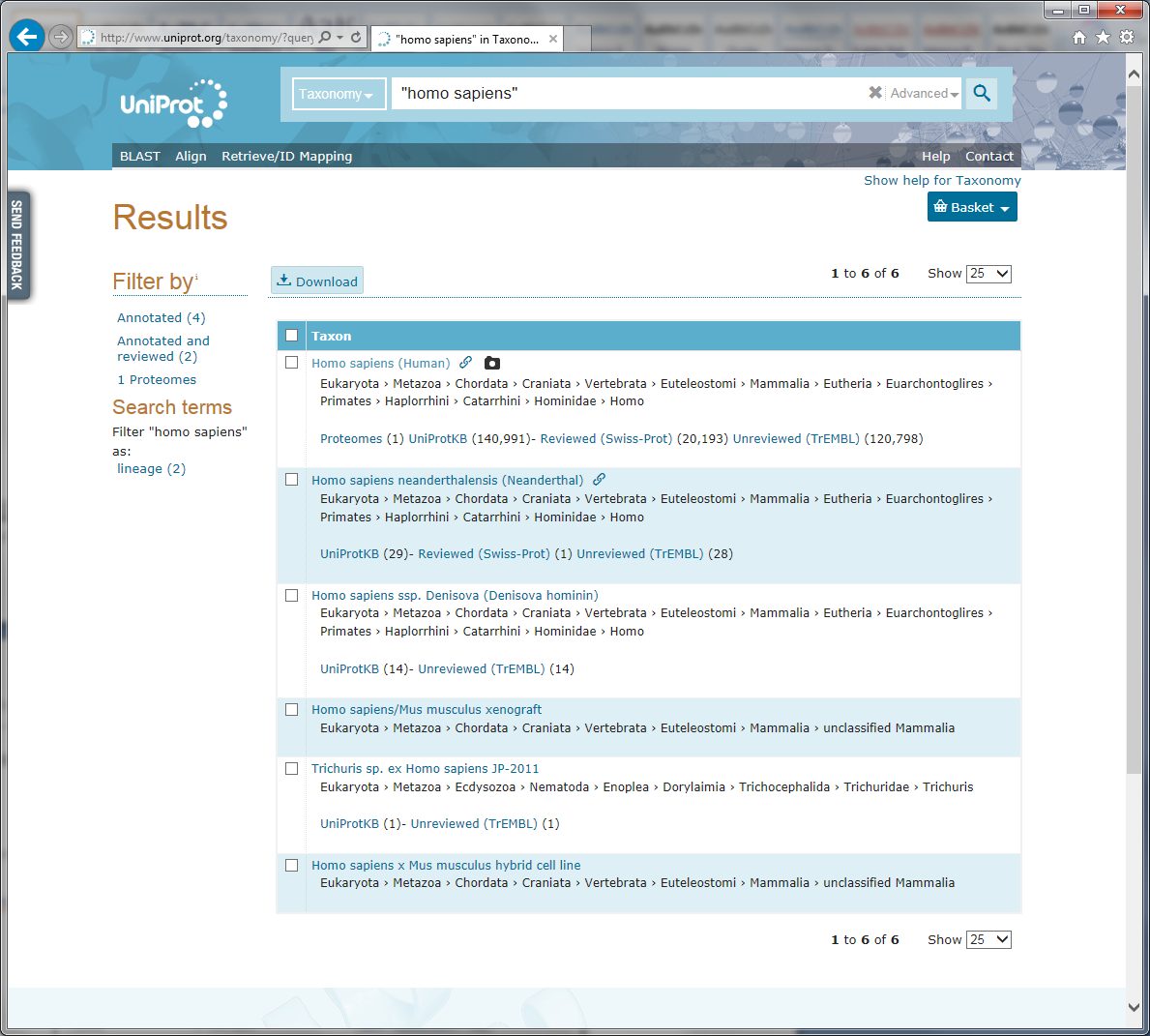
In order to identify peptides and proteins, we are going to compare the mass spectra to *in sillico* theoretic spectra deduced from a protein database. *Are there other database types that could be used to identify the spectra? Would it even be possible to identify the spectra without a database at all? [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[1](#_ENREF_5) database is a repository of choice for proteomics as it allies quality and quantity of protein sequences.

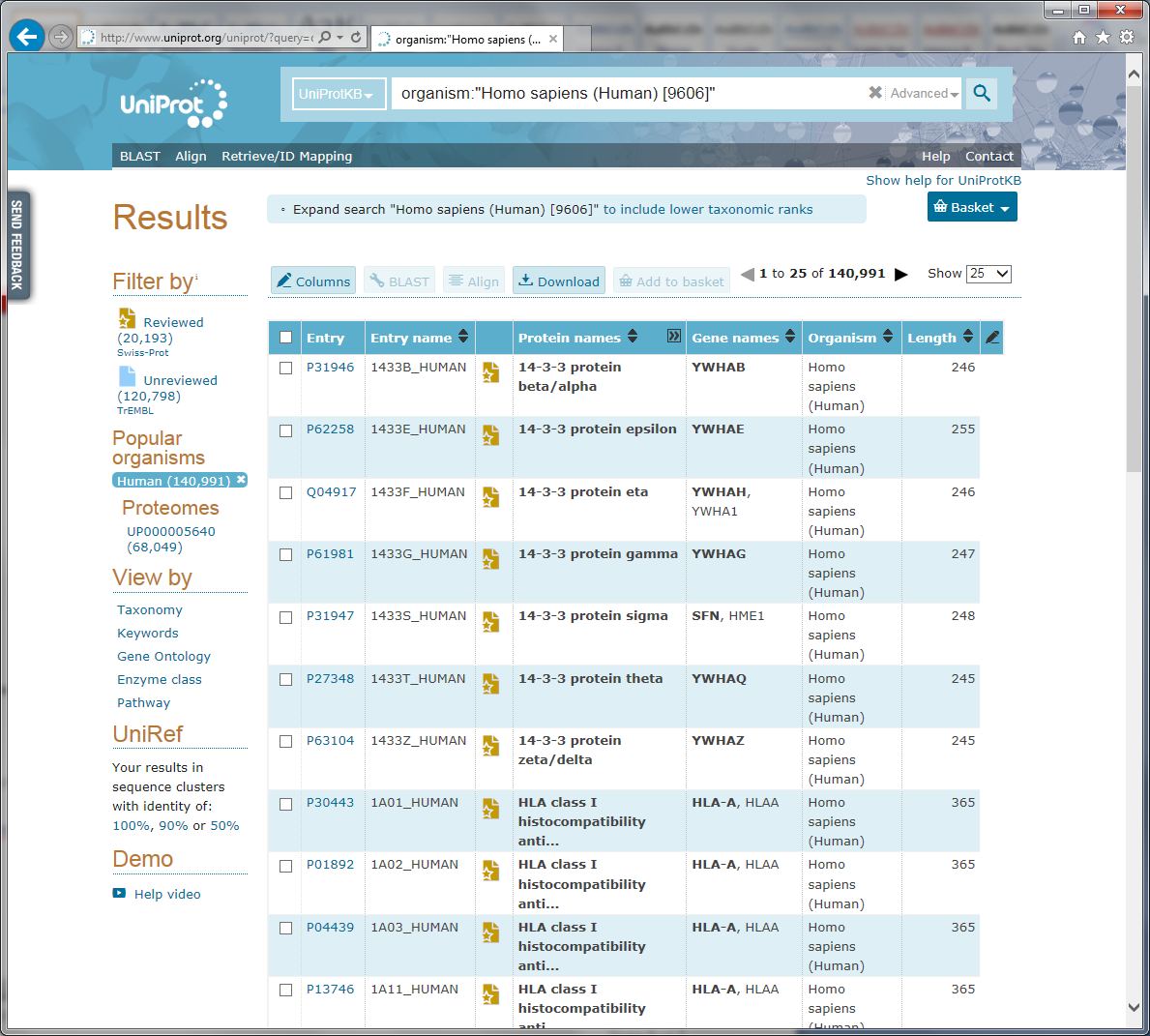
In order to optimize the database size, we will select only the species needed. The spectra in our example were obtained from a human sample. Go to the UniProt website (<www.uniprot.org>) and select “Taxonomy“ under “Supporting data” in the middle of the screen.



Next fill in “homo sapiens“ in the “Taxonomy“ field at the top and hit enter. UniProt retrieves six hits (as of September 2014). For the first result “Homo sapiens (Human)“ click the link “UniProtKB“.



This will show you all the human proteins in UniProt:



*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 ”Reviewed” (Swiss-Prot) option in the upper left corner to only list the reviewed proteins and then click the “Download“ button above the table. Make sure that “Download all“ is selected and the format is set to “FASTA (canonical)“. Click “Go“ to start downloading the protein sequences, and save the file as uniprot-human-reviewed-september-2014.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!*

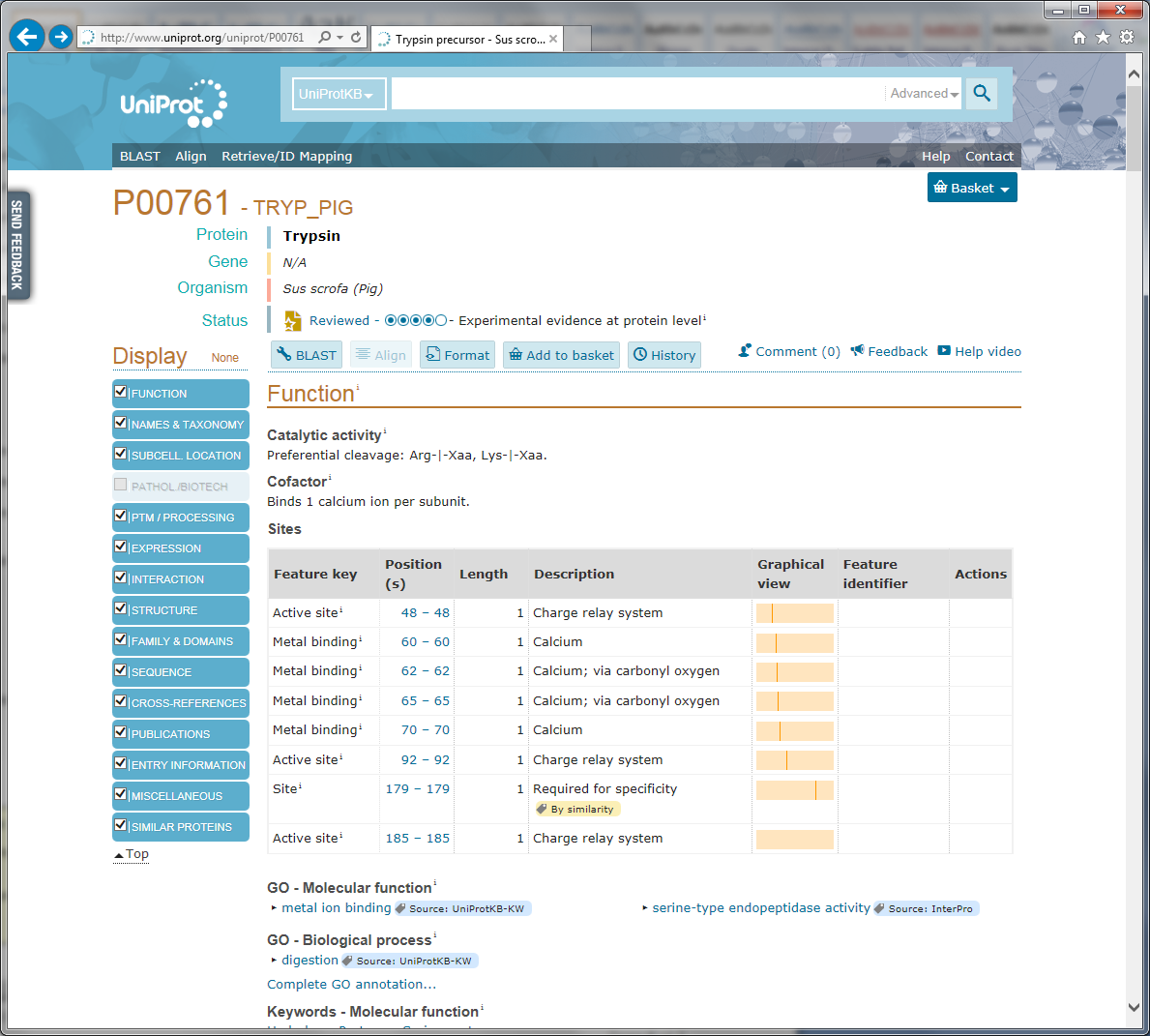
We now have a FASTA file with all the reviewed human protein sequences. However, our sample may also contain proteins from other species than the one we are studying, most often as a result of sample contamination.

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 sample2. A list of common contaminants can be found at the Global Proteome Machine3 (GPM) website (<http://www.thegpm.org/crap>).

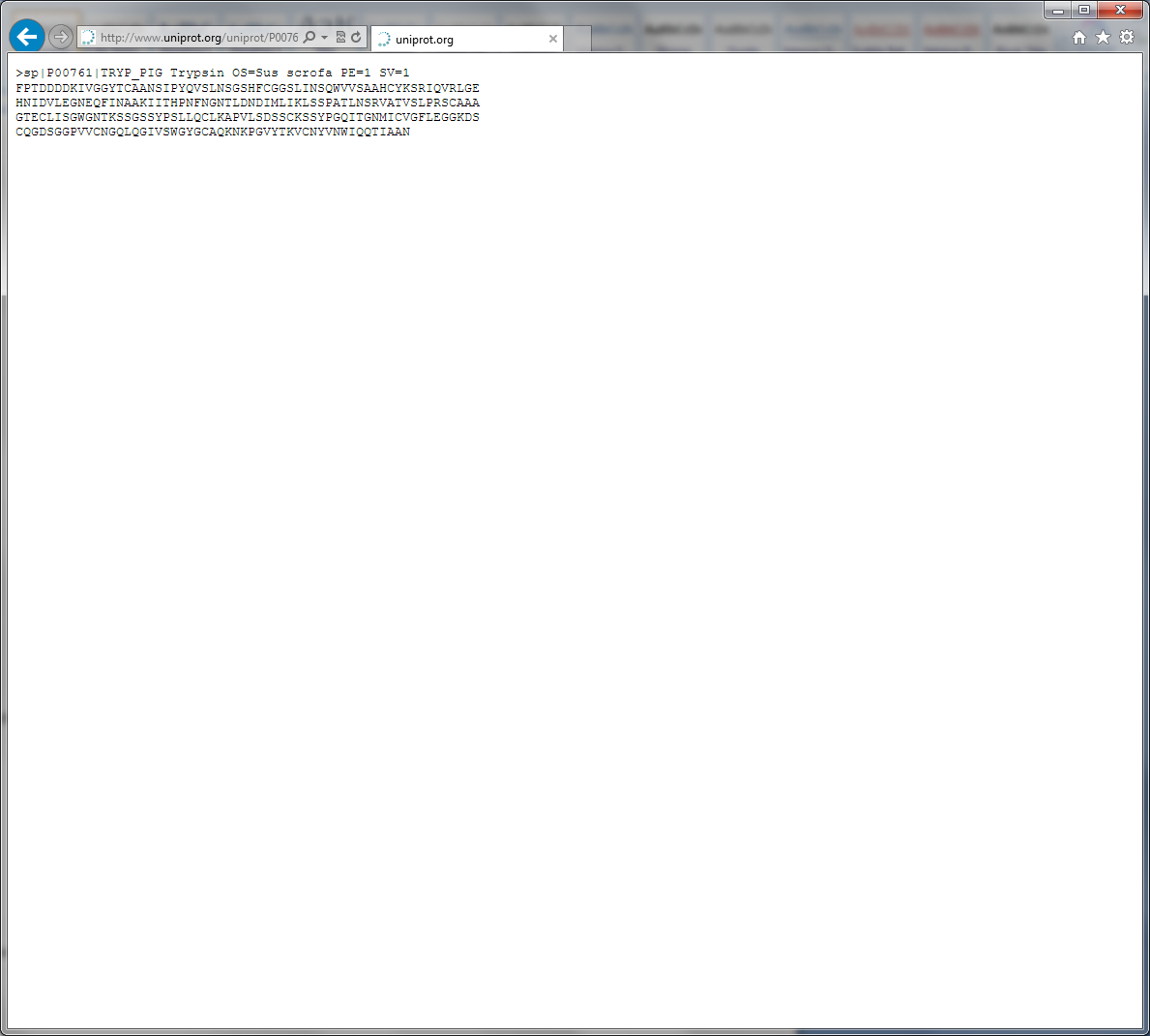
As we are here working with human data we do not have to add human keratin to our list of proteins. *Why not? [1.1d]*

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 webpage and search for the trypsin protein from pig: ”P00761”.



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



Open your uniprot-human-reviewed-september-2014.fasta file in a text editor like WordPad, 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. Save this new file as uniprot-human-reviewed-trypsin-september-2014.fasta.

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

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

MEANSGRLASELNHVQEVLEGYKKKYEEEVALRATAENEFVALKKDVDCAYLRKSDLEAN

VEALTQEIDFLRRLYEEEIRVLQSHISDTSVVVKMDNSRDLNMHCVITEIKAQYDDIATR

SRAEAESWYRSKCEEMKATVIRHGETLRRTKEEINELNRMIQRLTAEVENAKCQNSKLEA

AVAQSEQQGEAALSDARCKLAELEGALQKAKQDMACLIREYQEVMNSKLAWTLRSPPTGA

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://code.google.com/p/searchgui/wiki/DatabaseHelp>). 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).