Supplementary material and methods

Decoy database generation

The decoy database was generated using the peptide pseudo-reversed [1] method using the neXtProt 20150901 protein database. Instead of creating a database of entire decoy proteins, we altered each peptide sequence derived from the *in silico* digestion of the reference database using the proper enzymatic digestion rules (in our study the standard rules for trypsin). For each database entry, the peptides obtained were reversed and re-joined to calculate the corresponding entry in the decoy database. Using this approach, the decoy peptides matched the masses and the number of the target peptides considered by the search engine (Mascot 2.3). In house scripts in R/Bioconductor statistical environment [2] were used to implement the method.

False discovery rate (FDR) calculation

There are several methods reported in the literature to calculate the FDR using the target-decoy searching strategy [3]. There are basically two ways of performing the search using the selected search engine (Mascot) against these proteomic databases: using concatenated databases, when the decoy database and the target database are combined and searched together and using separately searches for the decoy and the target database. We used the second method, in which the FDR at PSM level is calculated as:

$$FDR = \frac{FP}{FP + TP} = \frac{D}{T}$$

Where D is the number of decoys passing a given threshold (*ion score*) that corresponds to the number of FP and T is the number of hits in target PSMs above that threshold (the sum of the number of false positives and true positives).

For the calculation of the protein FDR we used an in house implementation of MAYU software [4] that calculated the protein identification FDR on the basis of peptide identifications (PSM FDR < 1%). Using the HPP guidelines only those proteins with protFDR < 1% and with 2 or more proteotypic peptides were considered as positives identifications.

Estimation of the number of false positives identifications

The estimation of the FDR at the different levels of the analysis from the resulting set of PSMs, peptides and proteins was performed using the number of decoys identified as a measure of the number of false positives included in the selection:

$$FDR_{PSM} = \frac{\#PSM\ Decoy}{\#PSM\ Target}$$

$$FDR_{peptide} = \frac{\#Peptide\ Decoy}{\#Peptide\ Target}$$

$$FDR_{protein} = \frac{\#Protein\ Decoy}{\#Protein\ Target}$$

Estimation of the protein FDR using only proteotypic peptides

We incorporated a simpler strategy than the one described in MAYU to estimate the number of false positive proteins included in the selections performed using the bioinformatics workflow proposed in this study: PSM FDR < 1 % and intersection with the proteotypic peptides from the reference database. First, we processed the decoy database to obtain the set of decoy proteotypic peptides using the same methods applied to the target database. Then, we calculate the number of decoy proteins identified with decoy proteotypic peptides with an *ion score* higher than the minimum *ion score* in the target results. The estimated FDR of the selection was the ratio:

$$\widehat{FDR}_{protein} = \frac{\{\#decoy\ proteotypic\ proteins \mid ion\ score > \min(target\ ion\ score)\}}{\#target\ proteotypic\ proteins}$$

References

[1] Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol. 2010;604:55-71.

[2] Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (ed.), Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York, NY.

[3] Aggarwal S, Yadav AK. False Discovery Rate Estimation in Proteomics. Methods Mol Biol. 2016;1362:119-28.

[4] Lukas Reiter, Manfred Claassen, Sabine P. Schrimpf, Marko Jovanovic, Alexander Schmidt, Joachim M. Buhmann, Michael O. Hengartner, Ruedi Aebersold. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. Mol Cell Proteomics. 2009 Nov; 8(11): 2405–2417.