a) Schematic representation of the effect of searching very large datasets on the increase of the number of false protein identifications. Different proteins are represented by squares, peptides (in)correctly assigned through the PSM process are shown by stacked (red) green squares for each experiment and the cumulative view (ΣExp). A significant number of the proteins identified by a single peptide are false identifications, which tend to accumulate randomly scattered after repeated experimentation/runs. In contrast, true positive identifications tend to cluster to the proteins for which multiple matches are observed and accumulated. Recently, Mayu, a software tool to assess and control the protein FDR in large datasets has been published. Both Mayu (through estimation) and a study where all experimental single hit protein identifications above a give peptide probability threshold were manually validated have indicated that the percentage of false positives among single hits can be as high as 65%. However, verified single hit identifications tended to be shorter proteins and of higher predicted abundance (see Supplemental Figure 3).

b) Progression of the number of true positive versus overall number of distinct protein identifications as a function of the overall amount of peptide identifications above a chosen probability threshold. The data shown up to the horizontal line are experimental data from a dataset generated from the microbe Leptospira interrogans; beyond the horizontal line a simulation of the further progression is plotted assuming that the experimental set-up is unchanged. The difference between the green and black curve represent false positive protein identifications. Figure reproduced with permission from Ref. 3. These data suggest that the control of the false discovery rate at both, the peptide (PSM) level and protein inference level are critical when searching very large datasets.

1 Reiter, L. et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 8, 2405-17 (2009).