

## Supplementary Figure S5 Mascot evaluation

Mascot is one of the more popular probability based scoring schemes available. A limited investigation of Mascot v2.0 was carried out using the extended PSM dataset in tryptic search modes. Mascot was tested with a generic database corresponding to extended PSM dataset. The MS/MS ion search from Mascot was performed with the following options: mass accuracy 0.5 mass/charge, instrument type was chosen as ESI-Trap, parent mass tolerance was set at  $\pm 3.0$  Da (The same value used for Sequest and MASPIC in this paper) mass type was set to monoisotopic and the protein identification call was set at 'AUTO' mode. In the AUTO mode, Mascot determines the number of proteins present in the sample based on peptide matches. Additionally, for the tryptic search, the maximum missed cleavage was set at 5, (the same value used for Sequest in this paper). Mascot generates a summary report similar to DTASelect as an output. For evaluating Mascot, the MS/MS identifications in the summary report have to be classified into true and false identifications. The following rules were applied while generating the true and false identifications list: all peptides identified from protein sequences found in the standard mixture along with known contaminants are considered true identifications and the rest are considered false identifications, only the top-ranked peptides are retained, a peptide listed under multiple proteins will appear only once in either the true or false list and finally, the list of false identifications was further tested to make sure that peptides from known true proteins list are not present in it and vice-versa.

The values appearing under the column score in the Mascot summary report is called the Ion score. This is the primary score on the basis of which peptides are identified. Mascot also reports three additional scores for each identification. They are identity scores, homology scores and E value (column Expect in summary report). There are no peer-reviewed publications to address issues related to these additional scores, such as: what criteria and formulas are used to calculate these scores?, what percentage gain in identification is achieved with these scores?, and how to use these additional scores to evaluate Mascot. Our best efforts to evaluate Mascot with these additional scores showed little or no improvements. Moreover, Sequest and MASPIC have been evaluated purely based on their primary scores. Factors like number of candidates per

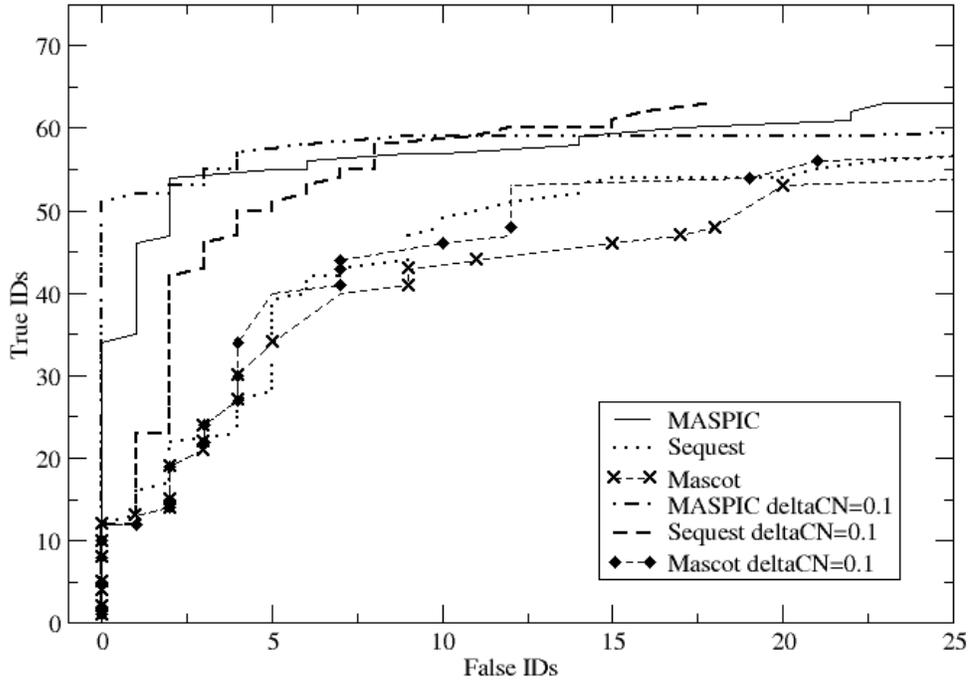
spectrum and the nature of distribution of false identifications on a spectrum by spectrum basis have not been considered.

Due to the above stated reasons, only the primary Ion scores reported by Mascot were used in generating the ROC curves. We computed  $\Delta CN$  for Mascot (from second best hit in “.dat” output file) and MASPIC in a similar to that of Sequest. The results of such an analysis have been shown in following figures. The figures show that among the three database systems evaluated, MASPIC still gives superior performance. The number of identifications clearing 95% confidence for Mascot was 394 ( $\Delta CN=0.0$ ) and 397 ( $\Delta CN=0.1$ ). The number of identifications clearing 95% confidence for Sequest was 371 ( $\Delta CN=0.0$ ) and 420 ( $\Delta CN=0.1$ ). The number of identifications clearing 95% confidence for MASPIC was 458 ( $\Delta CN=0.0$ ) and 473 ( $\Delta CN=0.1$ ). The above stated inspection and analysis also revealed an interesting aspect of Ion score. The Ion score is computed as  $-10\log_{10}(P)$  whereas MASPIC score is just the  $-\log_e P(\langle k_{ij} \rangle)$ . Unlike cross correlation and MASPIC score (see DTASelect.html files presented in the Supplementary website), the cutoffs for 95% confidence for Ion score generated by Mascot decreases with increasing charge state (38 for +1's, 25 for +2's and 17 for +3's on a tryptic search). This shows that the scores generated by MASPIC and Mascot, though being probabilistic negative log values, have different distributions based on charge state.

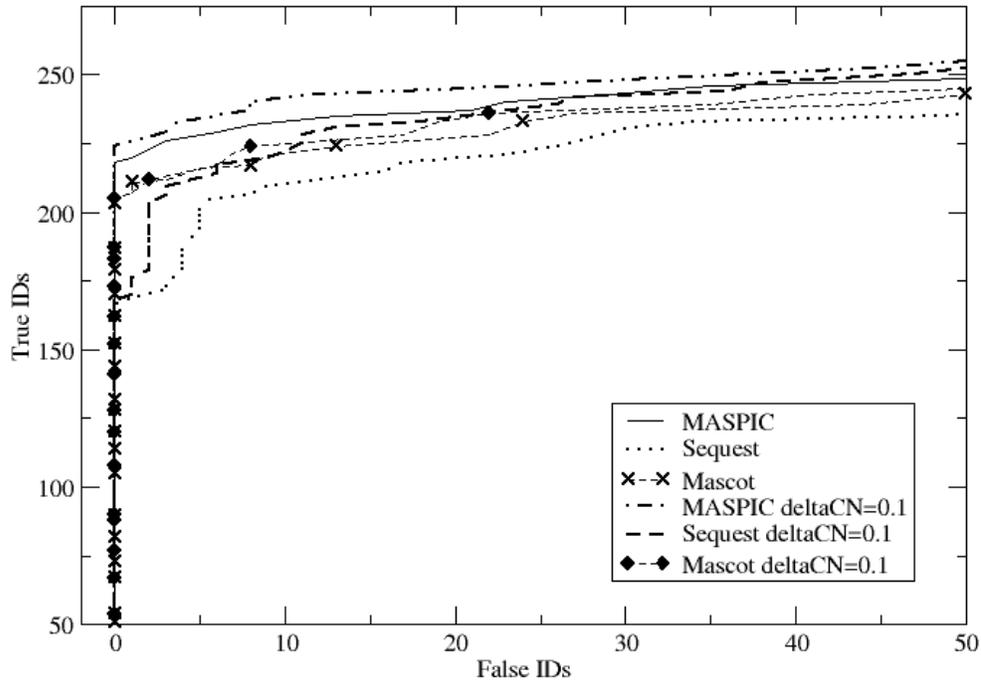
Figures S5: Tryptic search results for all three database search systems for the extended PSM. The ROC curves with  $\Delta CN=0.1$  were plotted after removing all the true and false identified spectra with  $\Delta CN < 0.1$ . Hence some of the curves can appear truncated.

(PTO)

ROC curves for Extended PSM  
(Parent charge 1, Extended PSM, Tryptic)



ROC curves for Extended PSM  
(Parent charge 2, Extended PSM, Tryptic)



# ROC curves of Extended PSM

(Parent charge 3, Extended PSM, Tryptic)

