Supplemental information 1

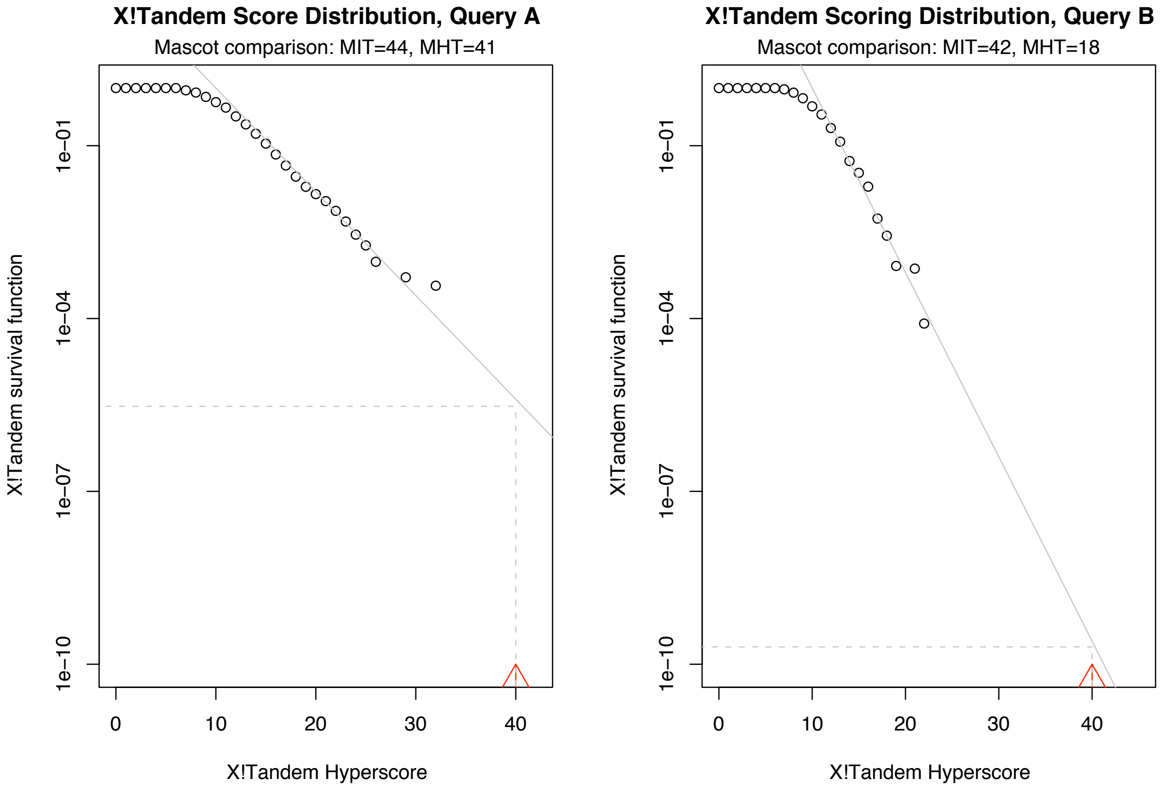

Figure 1: Exemplary survival functions from X!Tandem for two spectrum queries A and B. Although the number of peptide candidates for both queries is similar, there are apparent differences in the actual peptide score distributions. The survival functions were extrapolated for a score >= 40 that corresponds to a probability of approximately 3e-6 and 2e-10 for query A and B respectively. Given the number of peptides scored were 1e5, the expectation value of the former would be 0.3 while the expectation value of the latter would be 2e-5 (For a detailed explanation on how the survival function and expectation values are calculated, refer to Fenyo *et al.*[1]). Therefore, at a significance level of 0.05 the same score would have been considered highly significant for query B, but not for query A. In contrast, the MIT is inferred from the number of peptide candidates only, resulting in very similar thresholds of 44 and 42 for both queries. A hypothetical Mascot score of 40 would not have been considered significant for either query. On the other hand, the empirically derived MHT was 41 for query A and 18 for query B, thus classifying the peptide hit for query B as significant which agrees with the X!Tandem extrapolation example. It should be noted that the absolute scores and threshold values of X!Tandem and Mascot are not directly comparable.

Supplemental information 2

Gygi *et al.* proposed to exploit high accuracy MS data by searching at relaxed mass tolerance settings followed by mass accuracy filtering[2,3]. The rationale behind this is that the chance of finding a good peptide match in a relaxed mass window with many peptide candidates is greater than for a very stringent mass window with only a few peptide candidates. A correct and strong match is likely to remain the same, regardless of the size of the search space. On the other hand, it is more likely for a weak match arising from a poor spectra or from an incorrect peptide correlation to find a better alternative in a larger search space. A subsequent mass accuracy filtering step, which limits the matches to the experimental mass deviations, serves as useful discriminator between correct and incorrect matches. This is further illustrated in Figure 2.

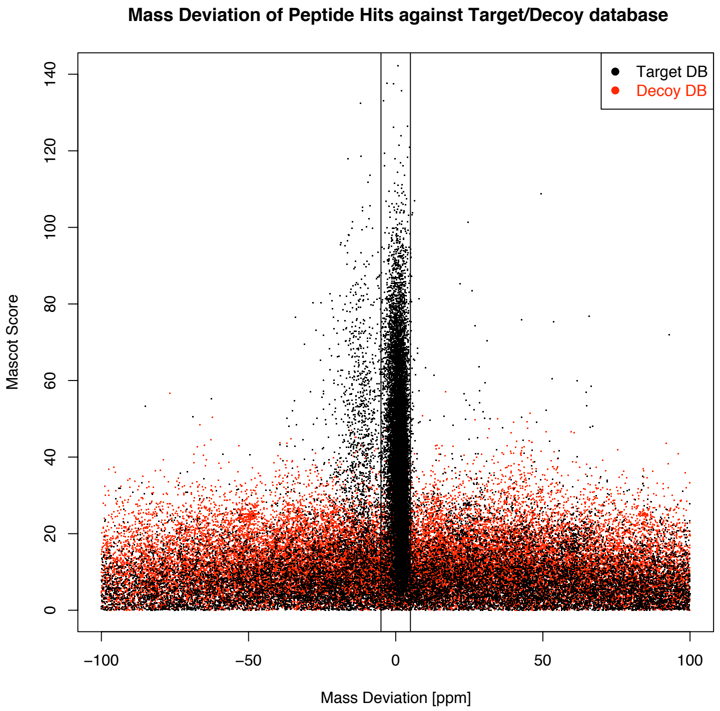

Figure 2: Distribution of all peptide matches obtained from a 1 Da MMD target and decoy database search showing the Mascot score and the mass deviations in ppm for a small window of ±100 ppm. Most mass deviations of high scoring peptide-spectra matches fell within the experimental mass errors that have been reported previously, 99% fell within ±20 ppm and 90% fell within ±5 ppm.

Supplemental information 3

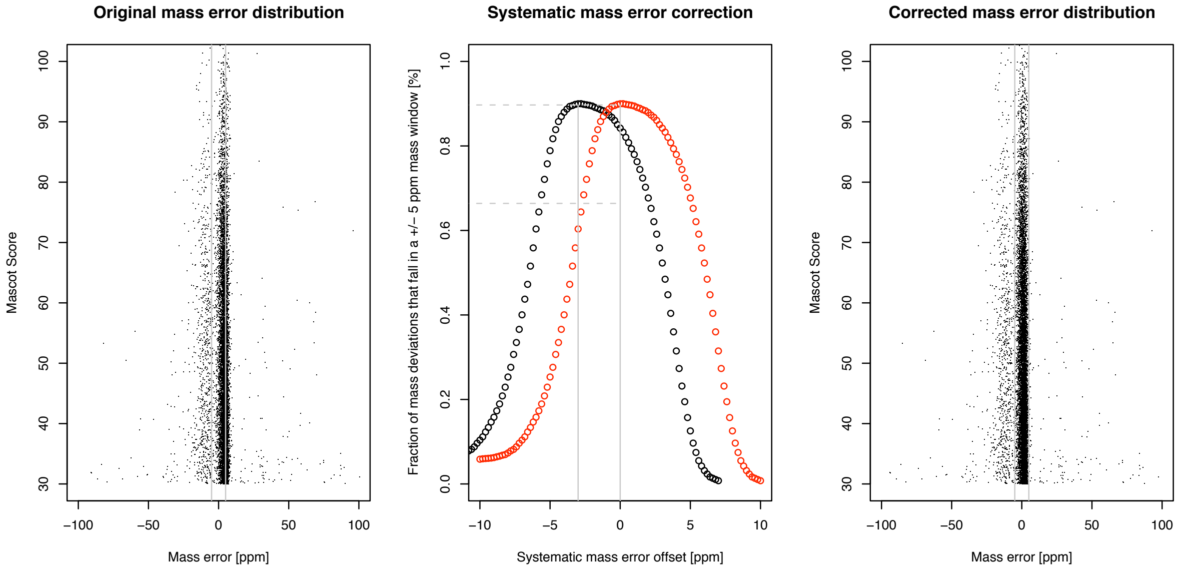

Figure 3: Mass error determination and correction of systematic mass errors. Left: the original mass deviations of all highly significant peptide matches. Centre: Systematic mass error correction that maximises the peptide assignments within a 5 ppm mass window. Right: After correction of the systematic mass error.

Supplemental information 4

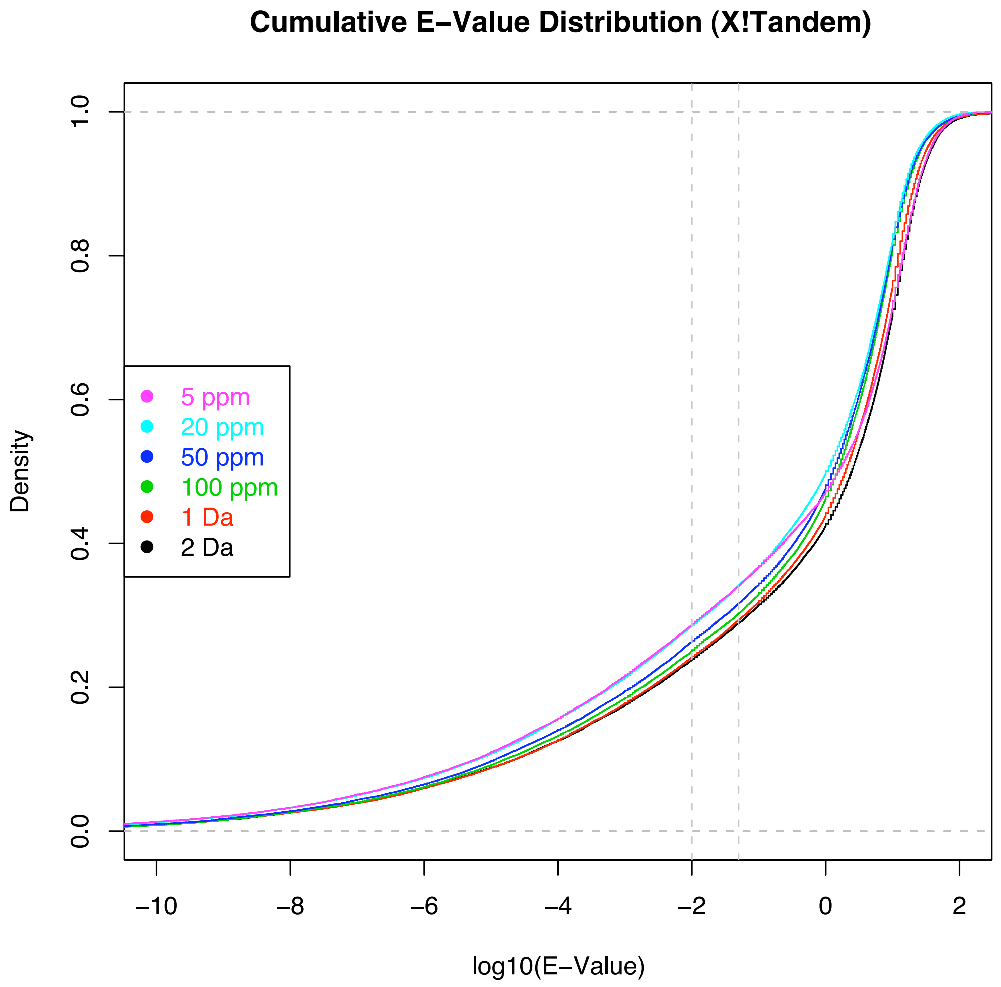

Figure 4: Spectra were searched with X!Tandem at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm MMD settings, while all other parameters were fixed. For each search the E-value distribution was drawn, indicating that the X!Tandem scoring is very robust over changes in search space. The E-values 0.01 and 0.05 are highlighted. The plot is in concordance with the ROC curve presented in the paper. Personal communication with Dr David Fenyo (The Rockefeller University) explained the robustness of the E-value distributions: Each E-value depends on the survival function and on the number of sequences scored (Fenyo *et al.*[1], equation 2). For X!Tandem in its current format, the term ”number of sequences scored” refers to the whole sequence database, regardless of the peptide mass tolerance setting and hence all variations seen in the E-values are the result of the slight differences in survival functions only.

Supplemental information 5a

Peak lists of sample 2 (8190 spectra) were searched with Mascot and X!Tandem against human IPI (June 2007, 68.322 sequences, 28.806.780 residues) including common external contaminants from cRAP (a maintained list of contaminants, laboratory proteins and protein standards provided through the Global Proteome Machine Organization, http://www.thegpm.org/crap/index.html). To minimise unexpected contaminants from the protein standard set[4], a very low concentration of 100 fmol was used. Parameters used: enzyme = trypsin; variable modifications = carbamidomethylation of cysteine, oxidation of methionine and deamidation of aspargine and glutamine; maximum missed cleavages = 2; peptide mass tolerance = 1 Da; product mass tolerance = 0.5 Da. A random and a reversed version of the sequence database was generated and searched under the same conditions.

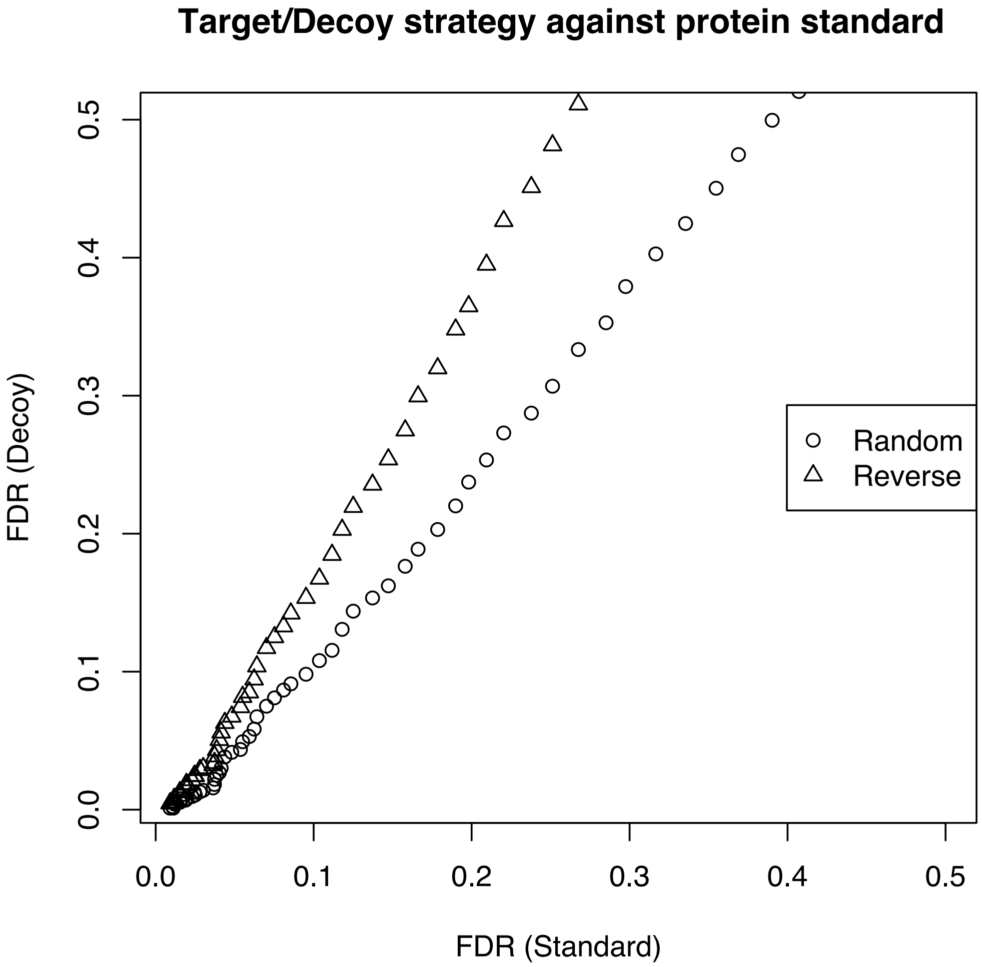

Figure 5: An experimental FDR, based on the known proteins of the set, can be determined as follows: any peptide hit that did not match against any of the 48 standard proteins or any of the external contaminants was considered a FP hit. The FDR rates were determined for a range of Mascot score cutoffs (10-50). Similarly, the estimated FDRs based on target/decoy searching were determined for both the randomised and reversed database. This enabled a comparison of actual FDRs with estimated FDRs (Figure 5). Both target/decoy searches show a linear relationship between the FDR determined by the protein standard and the target/decoy estimation, validating the target/decoy approach. However, the FDRs derived by the random database were closer to what was reported by the protein standard, which let us chose the random database as our decoy database for this study. Since these observations are based on relative comparisons, the same general findings would have been observed with any alternative decoy strategy.

Supplemental information 5b

To demonstrate the validity of our findings on an independent dataset, sample 2 was analysed by the same approach that was applied to sample 1. Data was first searched against a 50 ppm peptide mass tolerance to identify any systematic mass error (Figure 6).

The systematic mass error (-5 ppm) was corrected in the peaklist file and subsequently used to search at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm peptide mass tolerances. FDRs were determined and reported in Figure 7. The same trends as were seen for sample 1 were observed. For a more complete comparison, ROC curves were again compiled to enable a more comprehensive comparison (Figure 8).

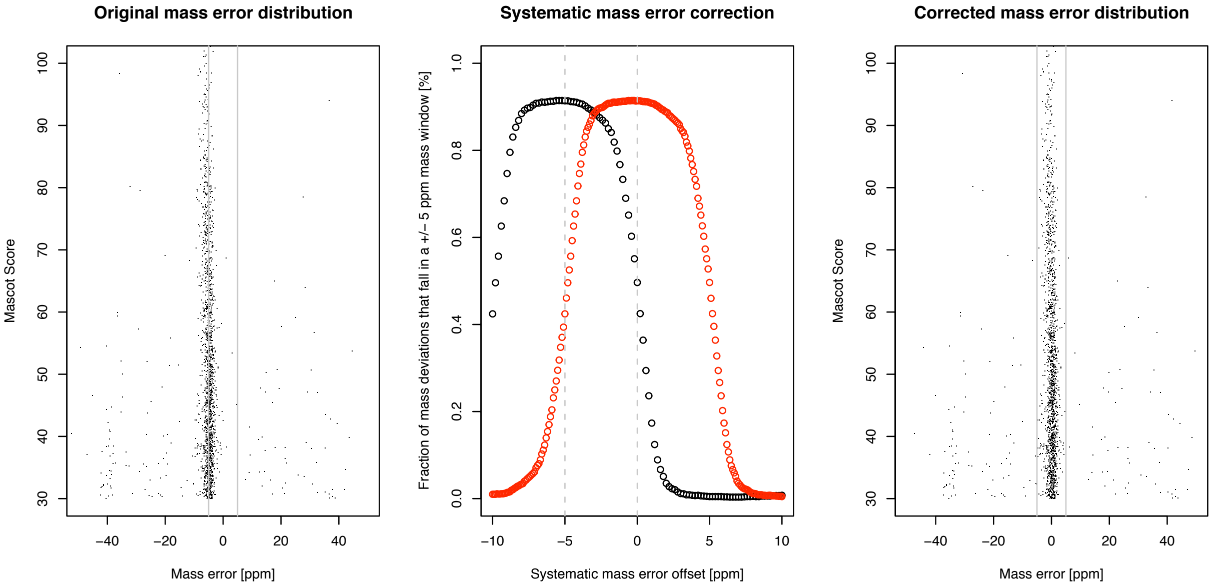

Figure 6: Systematic mass error correction (-5 ppm) for sample 2.

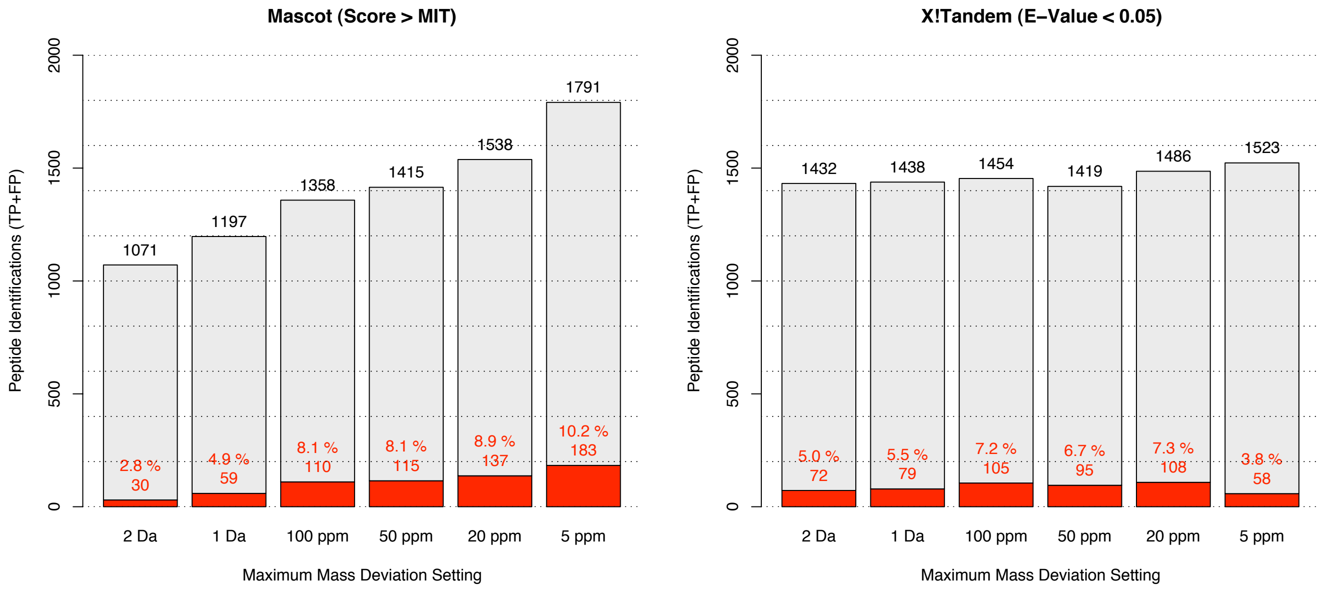

Figure 7: Comparative evaluation of Mascot and X!Tandem performance for sample 2. Mascot and X!Tandem searches were performed against a target and decoy database at different MMD settings. The total number of identifications is reported, the estimated number of true identifications is indicated in grey, while the estimated number of incorrect assignments is highlighted in red.

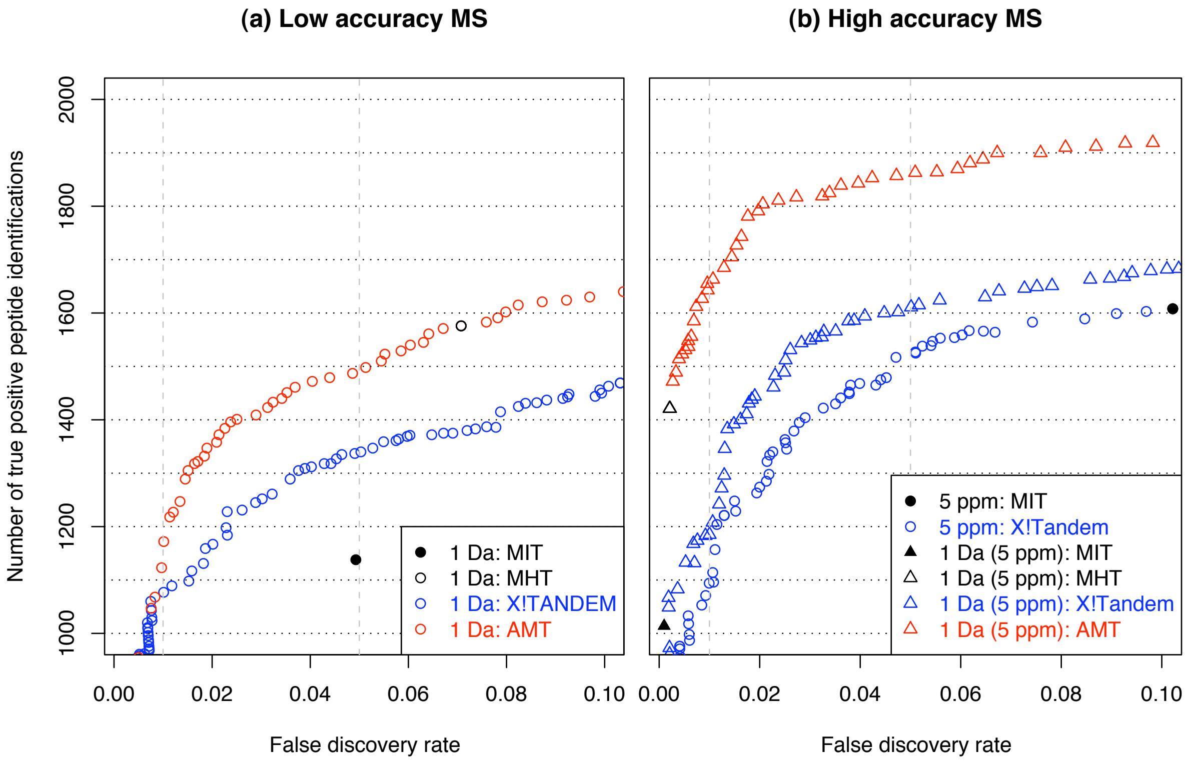

Figure 8: Comparison of the performance of X!Tandem and Mascot using the MIT, MHT and AMT for sample 2.

Supplemental information 6

RAW data, peak lists (with and without mass error correction) and Mascot results for both samples are available through ftp under the address: ftp://ftp.sanger.ac.uk/pub/mb8/mcp2008/

The Mascot search results have been included as tab delimited text files for manual investigations. They represent the original Mascot output from the 1 Da and 5 ppm search, supplemented with extra AMT related information where available (For the 5 ppm search the AMT was not applied for reasons explained in the paper. Further, peptide matches with no AMT classification have either not passed the threshold criteria or are not the top rank peptide match). For the 1 Da search, two columns (”1%_AMT” and ”5%_AMT”) were included that show whether a peptide was approved or not, applying the AMT at 1% and 5% FDR. The column ”Score >= MIT” indicates whether a peptide would have passed the Mascot Identity Threshold.

Further, for both the 1% and 5% FDR, the peptide hits obtained by the AMT were classified into 3 categories ”A”, ”B” and ”C”. Class ”A” represents peptide matches that passed the AMT with and without peptide mass filtering at 5 ppm. Just to recall, the mass filtering is applied on the 1 Da search prior to the AMT scoring to utilise the discriminative power of the high peptide mass accuracy. Class ”B” represents peptide matches, that were only identified using the AMT at 1 Da. Into this category fall all peptide matches that had a greater mass divergence than 5 ppm. Lastly, class ”C” shows all peptide matches that are solely approved for the mass filtered search at 5 ppm using its associated AMT. These are peptide matches with a rather low score, but due to the mass filtering discrimination, a much lower AMT threshold can be applied to achieve the same FDR.

Supplemental references

[1] Fenyo, D. and Beavis, R. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 75(4):768–774

[2] Everley, P., Bakalarski, C., Elias, J., Waghorne, C., Beausoleil, S., Gerber, S., Faherty, B., Zetter, B., and Gygi, S. (2006) Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy instrumentation. *J Proteome Res* 5(5):1224–1231

[3] Beausoleil, S., Villen, J., Gerber, S., Rush, J., and Gygi, S. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 24(10):1285–1292

[4] Klimek, J., Eddes, J., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P., Katz, J., Mallick, P. ., Lee, H., Schmidt, A., Ossola, R., Eng, J., Aebersold, R., and Martin, D. (2007) The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J Proteome Res*