Supplements 1, 2, 3 and 4

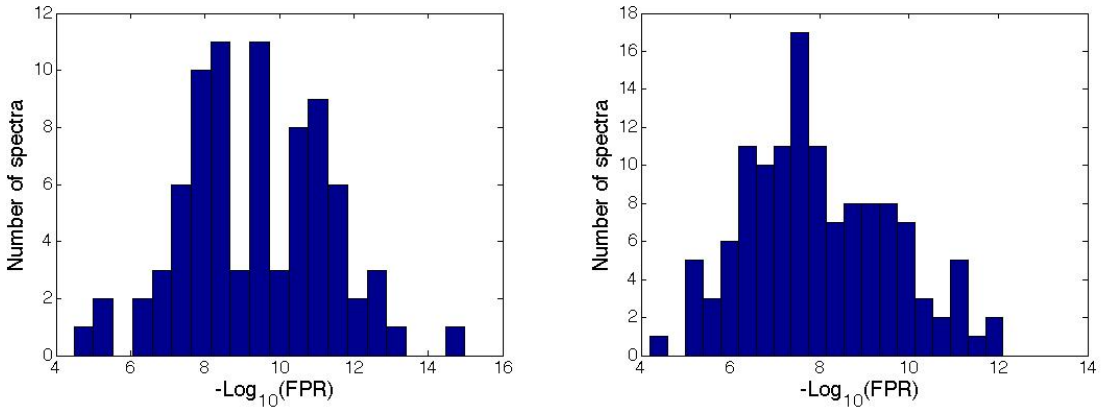
1 Critical analysis of estimates of statistical significance in existing MS/MS database search approaches

We found that the database search tools that we tested do not provide accurate estimates of the statistical significance of *individual* peptide identifications (they are often off by an order of magnitude). Fig.1(a) illustrates the analysis of X!Tandem identifications using the generating function approach. We selected all spectra of peptides of length 10 from the *Shewanella* dataset with X!Tandem E-value=0.01. Since all these spectral interpretations have the same E-value, they are expected to exhibit the same error rates in searches against decoy databases. However, the generating function analysis revealed that these spectral interpretations have vastly different FPRs as computed by MS-GF (orders of magnitude differences with large standard deviation). These remarkable variations raise a suspicion that either X!Tandem E-values are inaccurate, or MS-GF estimates are inaccurate. Below we show that MS-GF estimates are accurate thus raising concerns about the accuracy of X!Tandem E-values.

Recall that X!Tandem *estimates* FPR by constructing the empirical distribution of low scoring peptides for a given spectrum (either in decoy or in target database) and further fitting the tail of this distribution as described in [1]. The larger is the decoy database, the more accurate is the distribution, thus making it easier to fit it. Ideally, one can generate a giant decoy database with the goal to obtain the tail of the distribution explicitly instead of trying to fit it. Below we describe this experiment and demonstrate that fitting of the tail of the distribution leads to inaccurate estimates of FPR.

To compare the reported error rates with the probability of a spectrum matching a decoy database, we used *Shewanella-50000* dataset to construct 5 sets of spectra, each set containing 100 spectra with a fixed error rate (X!Tandem E-value), ranging from 10^{-1} to 10^{-5}. We then created a decoy database, 1000 times larger than the *Shewanella* protein database with 1.47 billion amino acids[1]. We searched the selected spectra again against this giant database with X!Tandem, and for each spectrum, counted the number of peptides that matched the spectrum with the same or better scores than the correct identification. For error rate of 0.05, we expect that the search on this 1000 times larger database will yield 50 peptide identifications, however we observe 4 identifications on average with high standard deviation (Figure 2(a)). Figure 2(b) shows similar distribution for InsPecT [2] search. InsPecT produces more hits (125 on average) than the expected number of 50 in the large decoy database. As shown in Table 1, the expected and the observed number of identifications differ significantly for all five datasets, thus indicating that the reported error rates may be unreliable.

[1]The search in such large database allows one to evaluate the FPR of *individual* spectra as opposed to the standard searches (with equally-sized target and decoy databases). Individual spectra typically have zero matches in the decoy database thus making it impossible to estimate the error rates of individual spectra.



(a) (b)

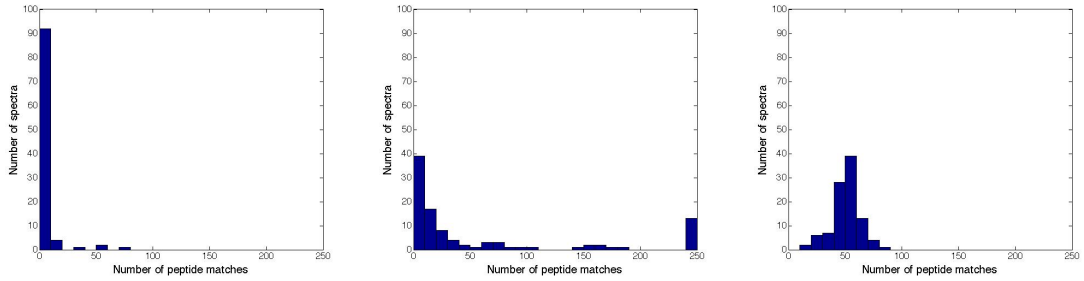
Figure 1: (a) Histogram of the FPRs (estimated by the generating function) of X!Tandem identifications for all spectra of length 10 (82 spectra) with E-value of 0.01. (b) Similar histogram for InsPecT identifications(128 spectra).

To compare these numbers against MS-GF, we generated high-scoring peptide reconstructions[2] for each spectrum such that their *spectral probability* is equal the p-value (divided by the size of the *Shewanella* database) under consideration, and determine how many of these reconstructions are found in the large database. For error rate of 0.05, Figure 2(c) shows that the observed number of hits in the database is close to the expected number of hits (50) for all spectra (compare with Figure 2(a,b). Table 1 shows that MS-GF computes accurate p-values for all five test datasets.

E-value	Expected Hits	Inspect Average	Inspect S.D.	X!Tandem Average	X!Tandem S.D.	MS-GF Average	MS-GF S.D.
0.1	100	205.4	328.6	12.4	24.3	105.6	27.3
0.05	50	125.5	268.6	4.0	11.4	51.6	12.2
0.01	10	77.5	210.3	1.4	9.6	10.0	4.4
0.001	1	7.8	28.1	0	0	1.08	1.1
0.00001	0.01	0.4	1.7	0	0	0.01	0.1

Table 1: Average and standard deviation (S.D.) of the number of peptide matches in a randomized decoy database of size 1000 times the size of the *Shewanella* database (by InsPecT and X!Tandem searches). A peptide match is reported while searching the spectrum S only if it has the same or better score than the original search of S in the *Shewanella* database. These numbers are compared with the expected number of peptide hits, for five different p-values reported by InsPecT/X!Tandem in the *Shewanella* database search. These numbers are contrasted with MS-GF, for which we generated top reconstructions for the given p-value and counted how many of these reconstructions were found in the database. In case of InsPecT we used p-values instead of E-values (InsPecT does not report E-values).

[2]Let $N_t(S)$ denotes the number of peptide reconstructions for spectrum S with score at least t and $Prob_t(S)$ denote the *spectra probability* of these reconstructions. Suppose we want to generate all high-scoring reconstructions with total probability 10^{-9}. The discrete nature of the scoring function may result in cases where $Prob_t(S) > 10^{-9}$ and $Prob_{t+1}(S) < 10^{-9}$ for some t. In such cases, we randomly select the appropriate number of reconstructions with score t such that their total probability sums up to 10^{-9}.



(a) (b) (c)

Figure 2: (a) Distribution of the number of peptide matches to a randomized decoy database of size 1000 times the size of the *Shewanella* database (X!Tandem search for 100 spectra with E-value 0.05). X-axis shows the number of peptide matches, and Y-axis shows the number of spectra that have these many matches. A peptide match is reported while searching the spectrum S only if it has the same or better score than the original search of S in the *Shewanella* database. (b) Similar figure for InsPecT search. Values larger than 250 on the X-axis were added to the histogram peak for value 250. (c) Similar figure for MS-GF, for which we generated the top reconstructions with the spectral probability 0.05 and counted how many of these reconstructions were found in the database. The expected number of hits in all searches is 50. InsPecT significantly overestimates the error rate, X!Tandem significantly underestimates the error rate, while MS-GF accurately computes the error rate (average number of peptide matches is 52).

2 Critical analysis of estimates of statistical significance in MS-GF

We conducted another experiment on a larger set of spectra to further validate that the generating function reports accurate FPR. For all spectra in the *Shewanella-50000* dataset, we generated high-scoring peptide reconstructions whose total probability sums up to 10^{-9}. We constructed ten random databases of size 10^7 aa and conducted exact string search for the peptide reconstructions in each of these databases. The expected number of hits in a database of this size is $10^7 \cdot 10^{-9} \cdot 50,000 = 500$. The observed number of spectra that have hits to the database was remarkably close to this number in all ten cases, with an average of 499.6 and standard deviation of 48. We repeated the experiment with p-value of 10^{-10}, and found 48.5 matches on average (standard deviation is 8.4), close to the expected 50 matches. This experiment demonstrates that MS-GF provides reliable estimates of the statistical significance of peptide identifications.

3 The spectral probability improves the sensitivity-specificity trade-off of SEQUEST/PeptideProphet database search

The spectral probability can be used to re-score the identifications obtained by existing database search tools. Below we show that it leads to improving the sensitivity-specificity trade-off of SEQUEST database search. We illustrate this result using *Shewanella-50000* dataset searched against the *Shewanella* database and the decoy database.

For each spectrum in the *Shewanella-50000* dataset, three different scores are used for analyzing the peptide identifications by SEQUEST and constructing ROC curves: (i) SEQUEST XCorr score, (ii) PeptideProphet [3] probability obtained from the SEQUEST scores (combined score) and (iii) spectral probability as reported by MS-GF for the SEQUEST identification.

For each score, a varying cutoff is used, and the number of spectra that have an identification with scores above the cutoff in the *Shewanella* database and the corresponding false discovery rate (ratio of the number of identifications on a decoy database (of same size) and the number of identifications in the *Shewanella* database) are plotted in Figure 3(a). As in the case of X!Tandem (see the main text), MS-GF results in larger number of identifications when compared to the SEQUEST XCorr scores, and even the combined PeptideProphet probabilities.

4 Performance of MS-GF for peptides of length 14

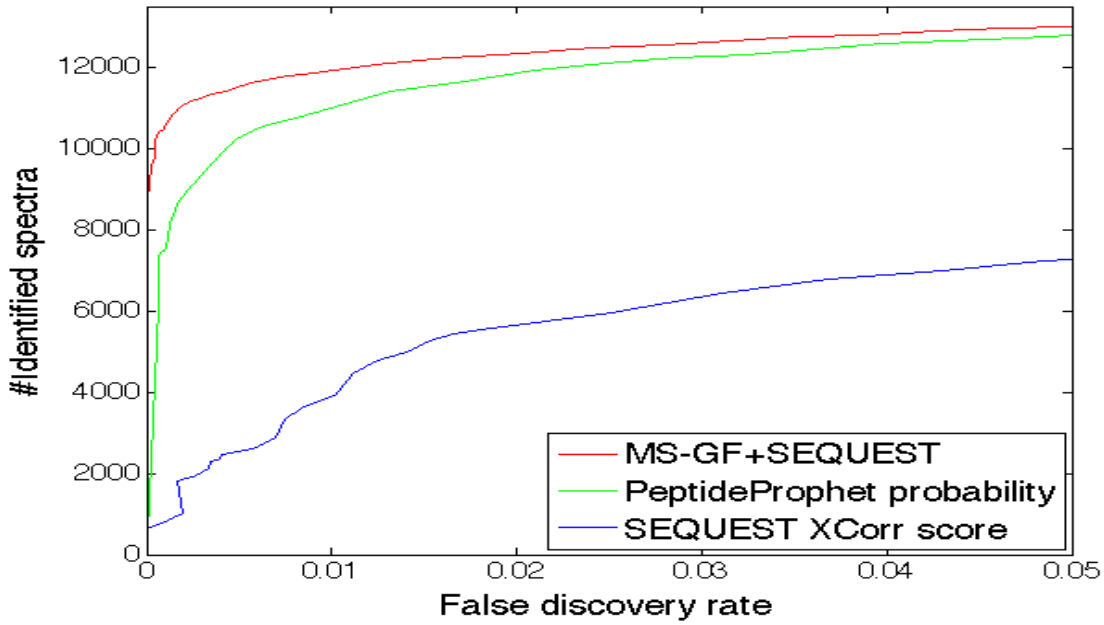
Below we analyze performance of MS-GF for peptides of length 14. Similar to Figure 4 in the main paper, we compare the sensitivity-specificity trade-off of MS-GF with X!Tandem on 50,000 randomly chosen spectra with parent mass range 1550-1605 Da (average peptide length ≈ 14 aa) from *Shewanella* dataset. Figure 4 here shows that MS-GF significantly improves the trade-off over X!Tandem raw and combined scores.

References

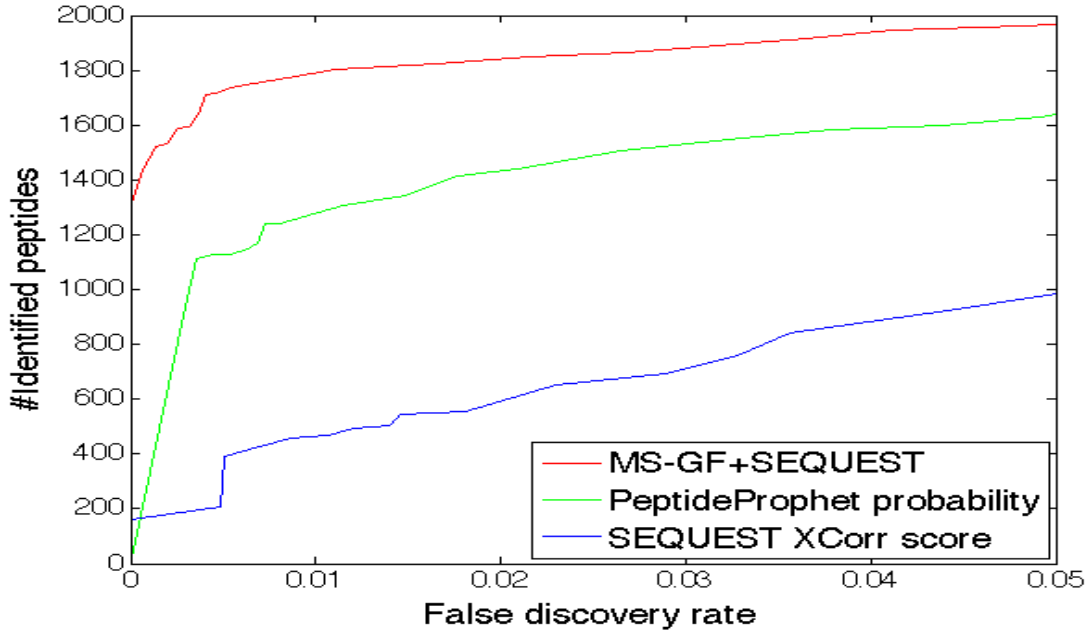
[1] Fenyo, D. & Beavis, R., A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem* **75**, 768–774 (2003).

[2] Craig, R. & Beavis, R., TANDEM: matching proteins with tandem mass-spectra. *Bioinformatics* **20**, 1466–1467 (2004).

[3] Keller, A., Nesvizhskii, A., Kolker, E., & Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383–5392 (2002).

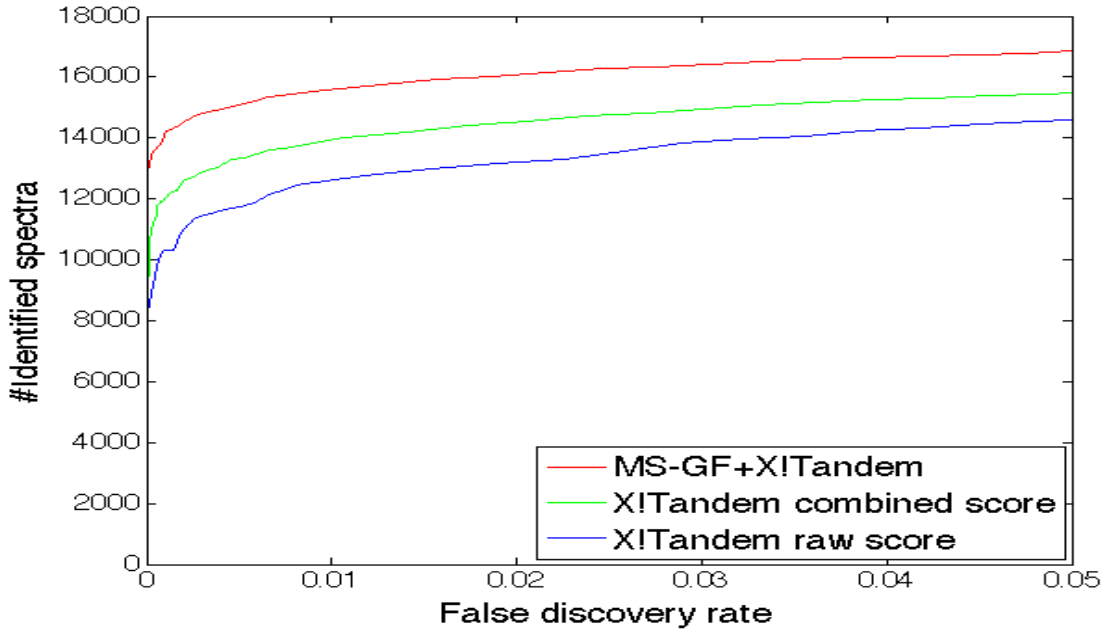


(a)

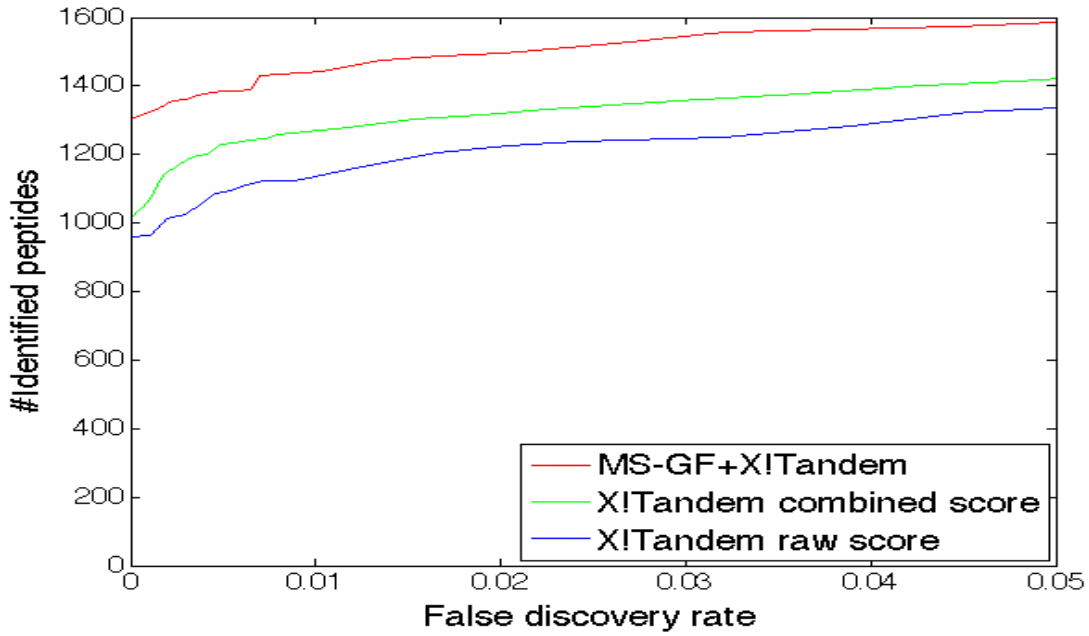


(b)

Figure 3: (a) MS-GF improves the performance of SEQUEST MS/MS database search. The number of spectra identified in the *Shewanella* database and the corresponding false discovery rate for different score cutoffs are reported. Three scores are compared (from top to bottom): (i) *MS-GF+SEQUEST*: spectral probability as reported by MS-GF for the SEQUEST identification, (ii) PeptideProphet probability obtained from SEQUEST scores and (iii) SEQUEST XCorr score. (b) Similar to (a), but counting the number of unique peptides identified in the *Shewanella* and the decoy database instead of the number of identified spectra.



(a)



(b)

Figure 4: Sensitivity-specificity trade-offs for spectra with parent mass range 1550-1605 Da (average peptide length is $\approx$ 14 aa).(a) MS-GF improves the performance of MS/MS database search X!Tandem when used independently and when used with spectral probabilities computed by MS-GF. The number of spectra identified in the *Shewanella* database and the corresponding false discovery rate for different score cutoffs are reported. Three scores are compared (from top to bottom): (i) *MS-GF+X!Tandem*: spectral probability as reported by MS-GF for the X!Tandem identification, (ii) *X!Tandem combined score*: X!Tandem E-value that uses the raw score as well as the distribution of scores of all peptides for the given spectrum and (iii) *X!Tandem raw score*: X!Tandem hypergeometric score. (b) Similar to (a), but counting the number of unique peptides identified in the *Shewanella* and the decoy database instead of the number of identified spectra.