SUPPLEMENTARY DATA

Supplementary Table 1. **Effects of X!Hunter preprocessing filters on LCQ search results.**

845 MS/MS from a curated dataset of LCQ spectra were evaluated by X!Hunter, using the same MS/MS for both the spectral library and the experimental dataset.

Filters used [a]	Matches [b]	Mismatches [b]	No match [c]	Spectra not evaluated [d]
8	734	56	1	54
5, 8	734	56	1	54
3, 8	734	56	1	54
2, 8	737	53	1	54
4, 8	774	7	9	55
1 – 8	777	3	10	55
1, 8	784	6	1	54
7, 8	785	5	1	54
6, 8	787	3	1	54
None	788	56	1	0
5	788	56	1	0
3	788	56	1	0
2	791	53	1	0
4	829	7	9	0
1 – 7	832	3	10	0
1	838	6	1	0
7	839	5	1	0
6	841	3	1	0

[a] Each number in this column refers to a filter named according to the subroutines in the X!Hunter code: (1) remove isotopes; (2) remove parent; (3) remove low mass ions; (4) dynamic range processing; (5) remove ions from parent neutral loss; (6) clean isotopes; (7) remove small ions (keep only number of ions required by user); (8) suppress noisy data.

[b] Match or mismatch to the known correct peptide sequences.

[c] No sequence assigned to the searched MS/MS spectra.

[d] Excluded from the search by X!Hunter.

Supplementary Table 2. Effects of X!Hunter preprocessing filters on LTQ search results.

Filters used [a]	Matches
8	109
5, 8	109
3, 8	109
None	111
5	111
3	111
2, 8	160
2	164
6, 8	385
6	387
1, 8	487
1	493
4, 8	693
4	700
7, 8	744
7	751
1 – 8	844
1 – 7	851

[a] Each preprocessing filter is numbered as described in Supplementary Table 1, footnote a.

The following tables are provided in separate Excel files, because of their size.

Supplementary Table 3. Large scale protein profile of Mascot search.

For a protein profile shown in this study, the first row is the name of the profiled dataset. The second, the third, and the fourth rows are the headers for each protein group. Each block starts with a group information row, follows varied number of protein information rows, ends with many peptides belonging to this group. The group data contains the group number (column B) and the number of peptides (column E). Each protein information row consists of the protein number (column B), the IPI accession number (column C), the molecular weight (column D), the number of peptides per primary (column E), percentage of amino acid covered (column F), and the protein description (column G). The peptide information has 5 columns: the protein number (column B), the peptide sequence (column C), the monoisotopic peptide mass (column D), the average peptide mass (column E), and the confidence (or consensus) level (column F). The following 9 columns are search dependent: the XCORR for Sequest, the Mowse for Mascot, the MAE SIM score generated by in-house application (Sun et al., 2007), the OMW, the standard deviation, the number of spectra assigned the same peptide sequence, and the previous number divided by charge 1 to 3. There are also many columns show the gene information from different resources for proteins. After the group information, there are summary data. For instance, numbers of proteins are supported by 1, 2, or more peptide sequences, and numbers of spectra are located in a protein.

Supplementary Table 4. Compare protein profile for large scale sample.

A compare profile has everything the same as a regular profile, but the first row which shows multiple names for a compare profile. In addition, the search dependent columns are repeated for multiple inputs.

Supplementary Table 5. Analysis of the protein overlaps between XM and SM

In this file, column A is the numbers of proteins found by SM and XM consensus. Column B is the number of cases for column A. For instance, the 4th row shows 1-1 for column A and 301 for column B. This means there are 301 cases that have the same peptide supporting the protein identification in both the Sequest and the X!Hunter results. Different charge forms of the same peptide are counted as different peptides, as described by Mann (Shi et al., 2007).

Supplementary Table 6. Peptide identifications for ABRF dataset.

This is a simplified MSPlus file. The first row shows the search name. The second row shows the names of columns. Sequest assigns a sequence to a DTA anyway, so there are no empty cells. For Mascot, empty cells for MSEQ1 (highest scoring Mascot sequence assignment) and Mowse1 means there was no result returned by the search program for that DTA. For X!Hunter, an empty cell means that this DTA was removed by the prefiltering. If the golden column shows "null" and the expectation value is 1000, no sequence assignment was made by the search program for that DTA.

Supplementary Table 7. Search results of a large scale dataset filtered by parent ion m/z mass accuracy, holding the peptide hit FDR at three constant values (0.5, 1, and 5), with additional restrictions on the peptide length, missed cleavage rules, charge. Comparisons are based on close FDRs, not exact. For 0.5 FDR, our data cannot go down so low for Mascot and Sequest only searches, so only the SM and XM consensus are shown. The numbers in each cell are the normal and decoy search results following by the FDR.

Search Method	Total Identifications	Unique Sequences	Protein Identifications
SM	6,239 / 33 (0.53)	3,072 / 26 (0.85)	846 / 25 (2.96)
XM	37,589 / 185 (0.49)	6,640 / 81 (1.22)	1,359 / 80 (5.89)

Search Method	Total Identifications	Unique Sequences	Protein Identifications
M	134 / 2 (1.49)	108 / 2 (1.85)	94 / 2 (2.13)
S	168 / 2 (1.19)	139 / 2 (1.44)	115 / 2 (1.74)
SM	44,527 / 444 (1.00)	7,249 / 200 (2.76)	1,449 / 197 (13.60)
XM	37,925 / 398 (1.05)	6,696 / 201 (3.00)	1,395 / 197 (14.12)

Search Method	Total Identifications	Unique Sequences	Protein Identifications
M	47,159 / 2,384 (5.06)	7,862 / 1,383 (17.59)	1,871 / 1,305 (69.75)
S	48,018 / 2,425 (5.05)	8,053 / 1,329 (16.50)	1,881 / 1,239 (65.87)
SM	45,443 / 2,291 (5.04)	7,593 / 1,205 (15.87)	1,738 / 1,138 (65.48)
XM	39,263 / 1,980 (5.04)	7,107 / 1,050 (14.77)	1,738 / 996 (57.31)

Supplementary Data. Algorithm of constructing ROC plot for a standard protein dataset.

```
INPUT: L: the set of DTAs
       g: the search program
       F: the known proteins
       N: number of spectra received sequences not in F
       P: number of spectra received peptide sequences in F
OUTPUT: R: a list of ROC points

S ← Sort L by g scores
FP ← 0
TP ← 0
R ← { (0, 0) }

for each d in S
  if g(d).assignedPepSeq is in F then
    TP ← TP + 1
  else
    FP ← FP + 1

  R.append(( FP/N , TP/P ))

Return R
```