SUPPLEMENTARY DATA

Supplementary Table 1. Comparison among high confidence assignments.

This is a very large table and is included as a separate pdf file.

Supplementary Table 2. Total frequency of amino acids.

Amino Acid	Frequency [a]	Amino Acid	Frequency [a]
A	8.3%	M	1.8%
C	0.5%	N	4.1%
D	6.4%	P	6.4%
E	9.1%	Q	5.1%
F	3.2%	R	3.0%
G	7.4%	S	7.1%
H	2.8%	T	5.5%
I	4.9%	V	6.6%
K	5.3%	W	0.8%
L	9.3%	Y	2.3%

[a] Frequencies calculated as total amino acids in 11,849 high confidence peptide sequence assignments, divided by the sum of all amino acids.

Supplementary Table 3. Data from high-confidence assignments of the test dataset used to determine SCX rules.

		Number of basic residues			
SCX fraction number		**1**	**2**	**3**	**≥ 4**
SCX #5	High-confidence assignments	16	0	0	0
	Percentage of spectra	100	0	0	0
	Acceptance rule	Yes			
SCX #6	High-confidence assignments	31	1	0	0
	Percentage of spectra	96.6	3.1	0	0
	Acceptance rule	Yes			
SCX #7	High-confidence assignments	26	1	0	0
	Percentage of spectra	96.3	3.7	0	0
	Acceptance rule	Yes			
SCX #8	High-confidence assignments	42	0	0	0
	Percentage of spectra	100	0	0	0
	Acceptance rule	Yes			
SCX #9	High-confidence assignments	43	0	0	0
	Percentage of spectra	100	0	0	0
	Acceptance rule	Yes			
SCX #10	High-confidence assignments	35	0	0	0
	Percentage of spectra	100	0	0	0
	Acceptance rule	Yes			
SCX #11	High-confidence assignments	41	0	0	0
	Percentage of spectra	100	0	0	0
	Acceptance rule	Yes			
SCX #12	High-confidence assignments	28	13	0	0
	Percentage of spectra	68.3	31.7	0	0
	Acceptance rule	Yes	Yes		
SCX #13	High-confidence assignments	6	27	0	0
	Percentage of spectra	18.2	81.8	0	0
	Acceptance rule	Yes	Yes		
SCX #14	High-confidence assignments	2	56	0	1
	Percentage of spectra	3.4	94.9	0	1.7
	Acceptance rule		Yes		
SCX #15	High-confidence assignments	0	60	0	0
	Percentage of spectra	0	100	0	0
	Acceptance rule		Yes		

		Number of basic residues			
SCX fraction number		**1**	**2**	**3**	**≥ 4**
SCX #16	High-confidence assignments	0	63	1	0
	Percentage of spectra	0	98.4	1.6	0
	Acceptance rule		Yes		
SCX #17	High-confidence assignments	0	65	3	0
	Percentage of spectra	0	95.6	4.4	0
	Acceptance rule		Yes		
SCX #18	High-confidence assignments	0	70	9	0
	Percentage of spectra	0	88.6	11.4	0
	Acceptance rule		Yes	Yes	
SCX #19	High-confidence assignments	1	64	20	0
	Percentage of spectra	1.2	75.3	23.5	0
	Acceptance rule		Yes	Yes	
SCX #20	High-confidence assignments	0	42	15	0
	Percentage of spectra	0	73.7	26.3	0
	Acceptance rule		Yes	Yes	
SCX #21	High-confidence assignments	0	29	33	0
	Percentage of spectra	0	46.8	53.2	0
	Acceptance rule		Yes	Yes	

Supplementary Table 4. Protein profile generated by IsoformResolver.

This is a very large table and is included as a separate Microsoft Excel file.

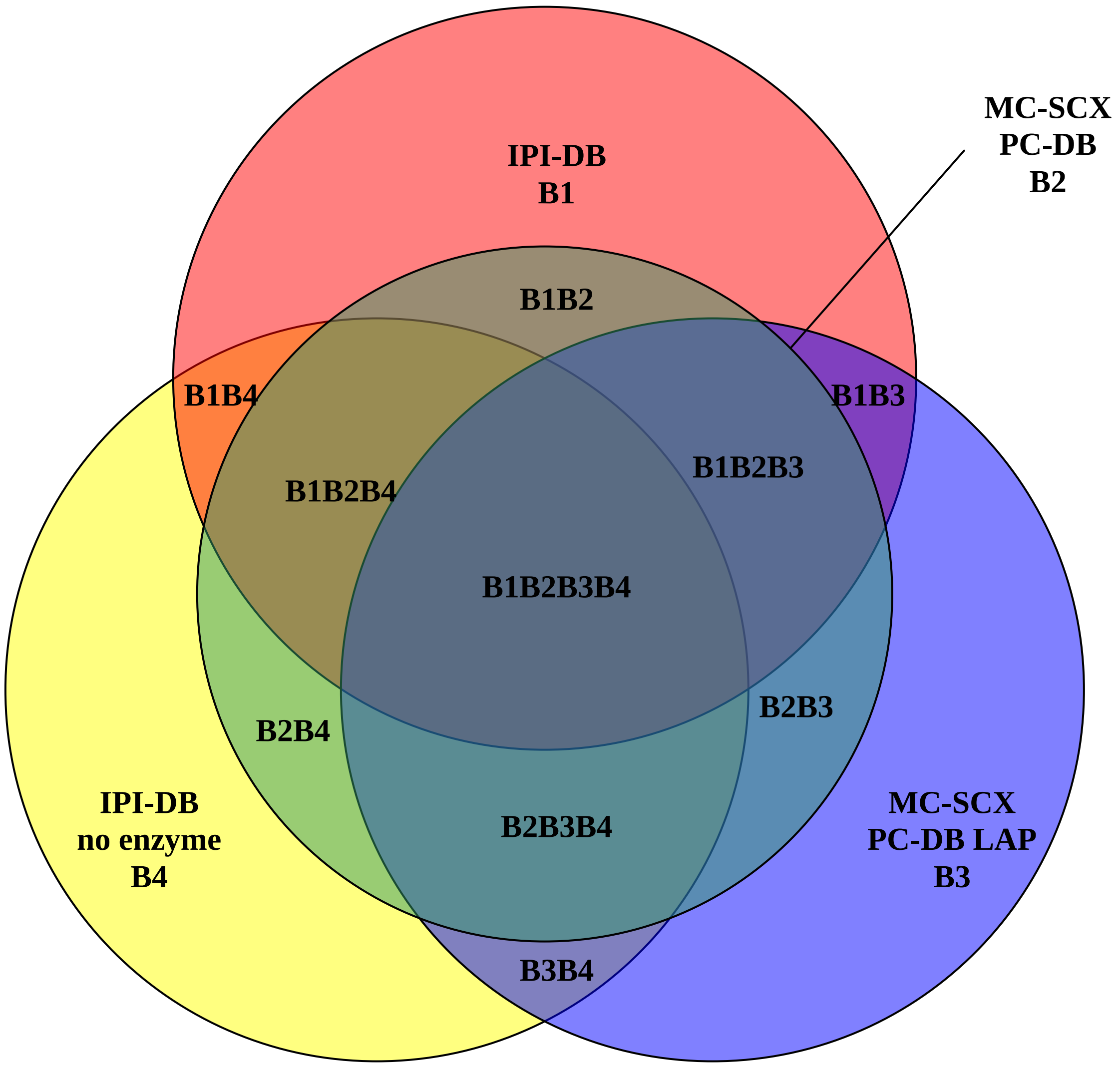
Proteins are grouped together IsoformResolver based on presence of common peptides. Briefly, it assembles the peptide sequences that MSPlus has validated, calls the information about the peptide from the database, and assembles all information regarding a peptide from multiple search results. There are 6 different results compared in this table. It determines whether a peptide is unique to a protein entry or found in more than one entry, classifying the information to provide a minimum protein list that accounts for all the peptides, but displays alternatives for the user to survey. It also compiles the information about the highest score found for each peptide sequence, the number of times a peptide sequence is observed, and the ion charge distribution, and calls information from the PC-DB to obtain IPI accession numbers, protein MW, gene name, GO codes, and accession numbers to RefSeq, ProFam, and other informatics database entries. IsoformResolver treats as identical all peptide isoforms that cannot be distinguished given the mass accuracy of the LCQ (e.g., replacement of I for L), although the presence of these alternatives in the dataset is noted in the output. However, in this table, this function is disabled in order to see the differences. There are 4 header lines in this table. The first line includes the names of different comparing results. The second line indicates the fields of the first line for each given group. The third line is the headers of the protein information in each protein group. Those protein information starts from the second line and ends before the peptide information in each protein group. The fourth line is the headers of the peptide hit information.

SUPPLEMENTARY FIGURE LEGENDS

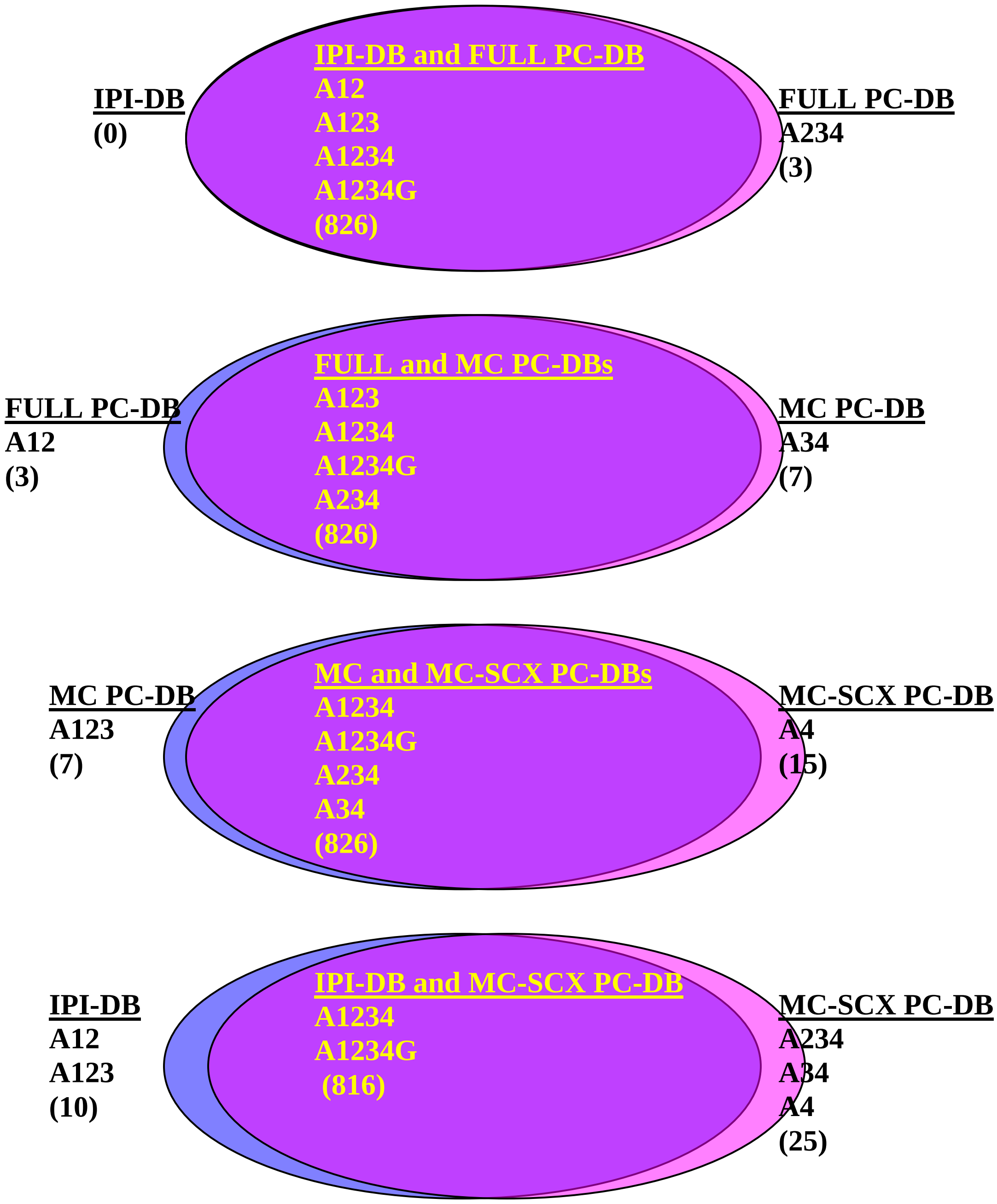
Supplementary Figure 1. Venn diagrams showing the relationships between the classes of DTA files shown in Supplementary Table 1 and summarized in main text Table 4. Each set of data unique for the IPI-DB or PC-DBs is given a unique designation where B1, B2, B3, and B4 indicate the assignment unique to IPI-DB, MC-SCX PC-DB, MC-SCX LAP PC-DB, and IPI-DB no enzyme, respectively. Other classes represent combinations of databases as shown in panel A. (A) the Venn diagram is shown for classes in Supplementary Table 1. In order to represent the overlap between 4 databases, one of them is represented as the middle circle. Because the unique class for B2 cannot be represented in the Venn, it is indicated at the upper right hand corner. (B, C) Venn diagrams are shown for the comparison between databases shown in main text Table 4B. For example, the set of DTAs compared between IPI-DB and FULL PC-DB can be related to Supplementary Table 1 as follows: DTA files unique to the FULL PC-DB search are found in section A234 in Supplementary Table 1A; no files unique to the IPI-DB search were observed; DTA files observed in both searches are found by combining sections A12, A123, A1234, and A1234G. Panel B and C represent tryptic and no enzyme searches, respectively.

Supplementary Figure 2. Reducing database size improves discrimination between correct *vs* false positive assignments for both MH_2^{+2} and MH_3^{+3} ions. Plots shown in main text Fig. 3 were subdivided according to charge. Due to the lower limit on peptide size (≥ 950 Da, ≥ 9 aa), the number of singly charged ions were insufficient to generate a smooth distribution and are not shown. Both doubly and triply charged ions show similar shifts in distribution to lower scores upon comparing protein database searches against PC-DB searches, as measured by (A, B) Sequest and (D, E) Mascot results. (E, F, G, H) Results of searching the dataset subsets shown in panels A, B, C, D respectively, using inverted databases. For XCorr score distributions of incorrect assignments, the triply charged ions show a narrower distribution compared to doubly charged ions, primarily due to the shift to higher scores on the lower end of the distribution. Thus, the low scoring incorrect assignments have an overall higher charge distribution for triply charged vs doubly charged ions, but the high confidence threshold is unchanged.

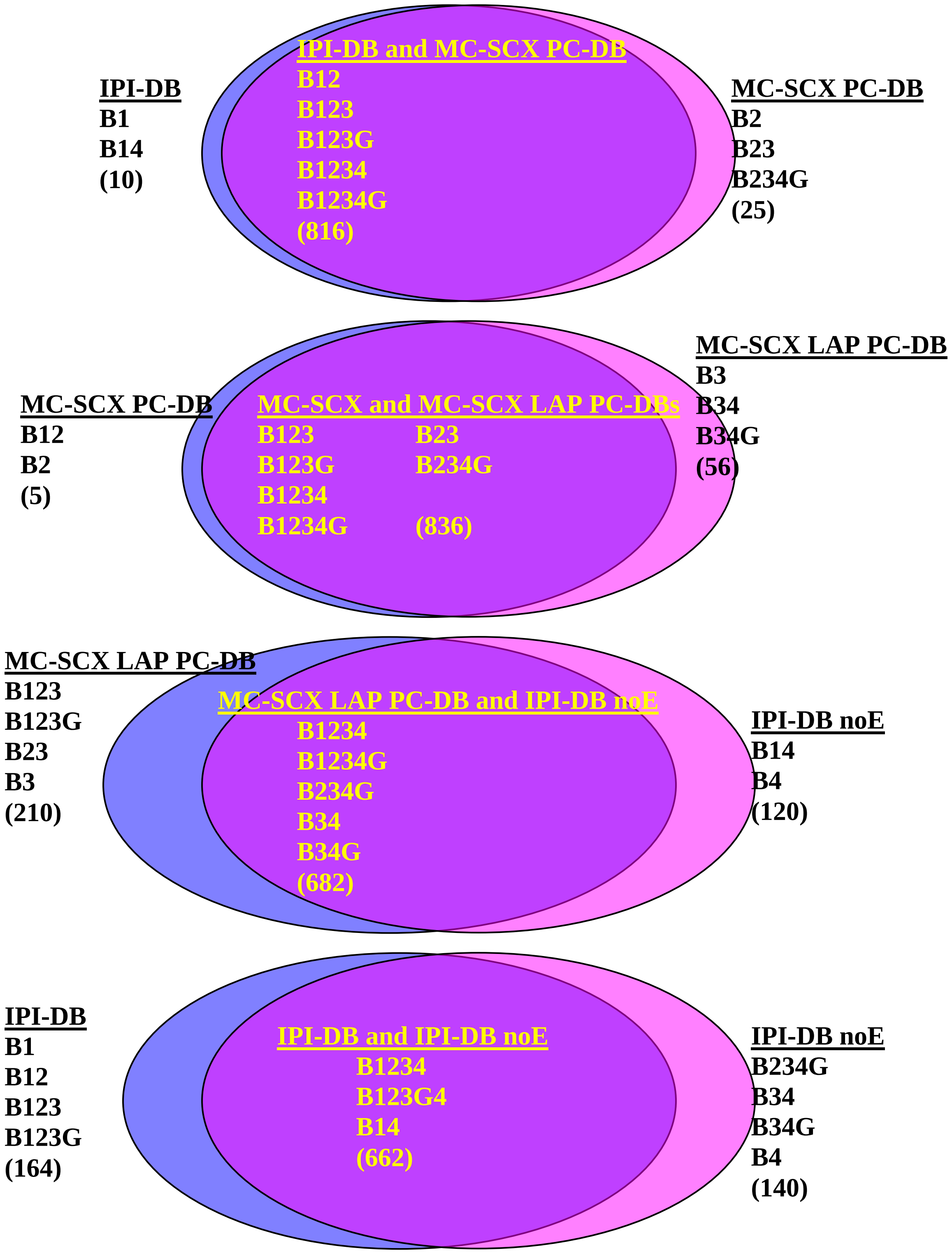
Supplementary Figure 1A. Over all view of comparison



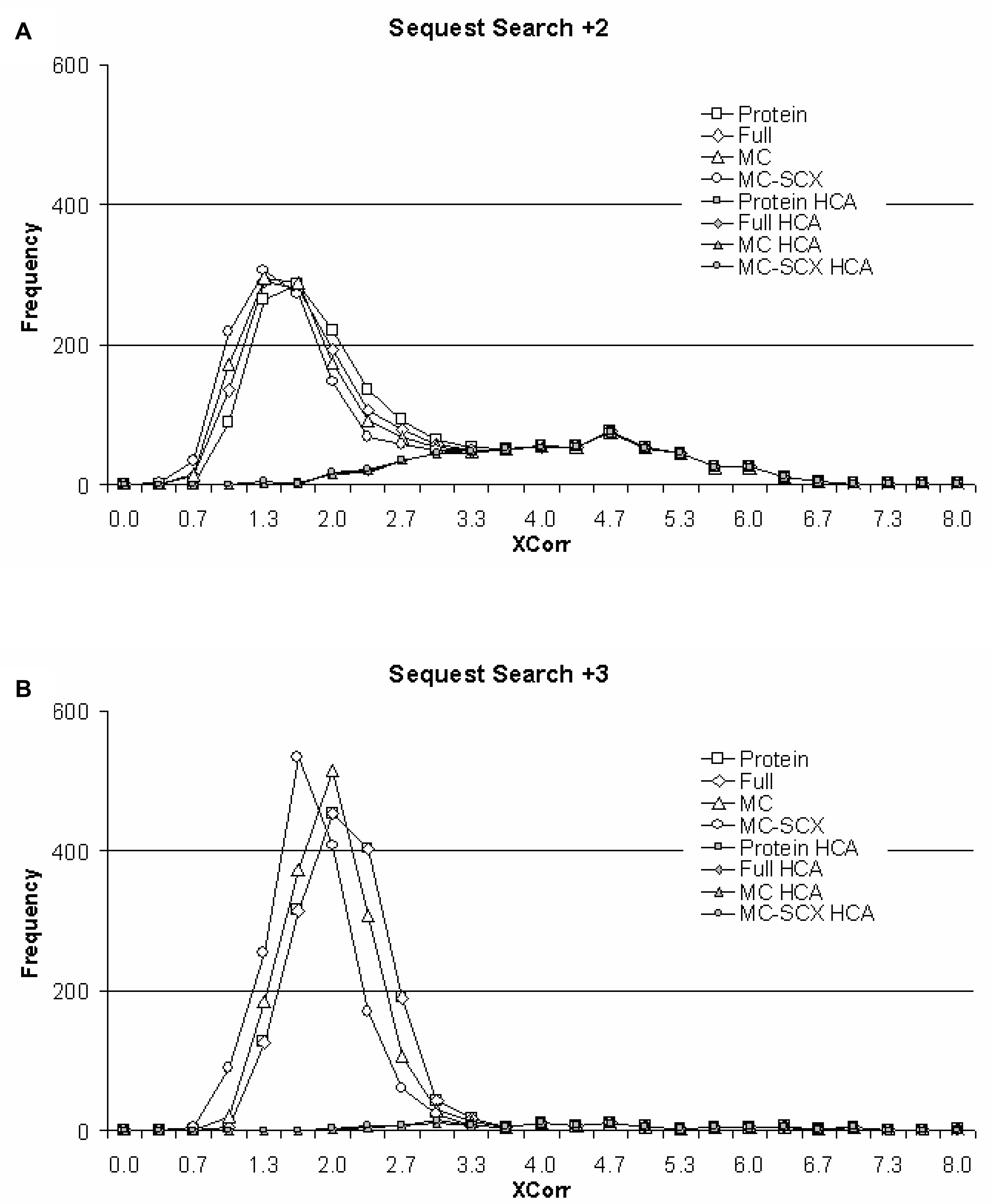
Supplementary Figure 1B. Comparison of tryptic results

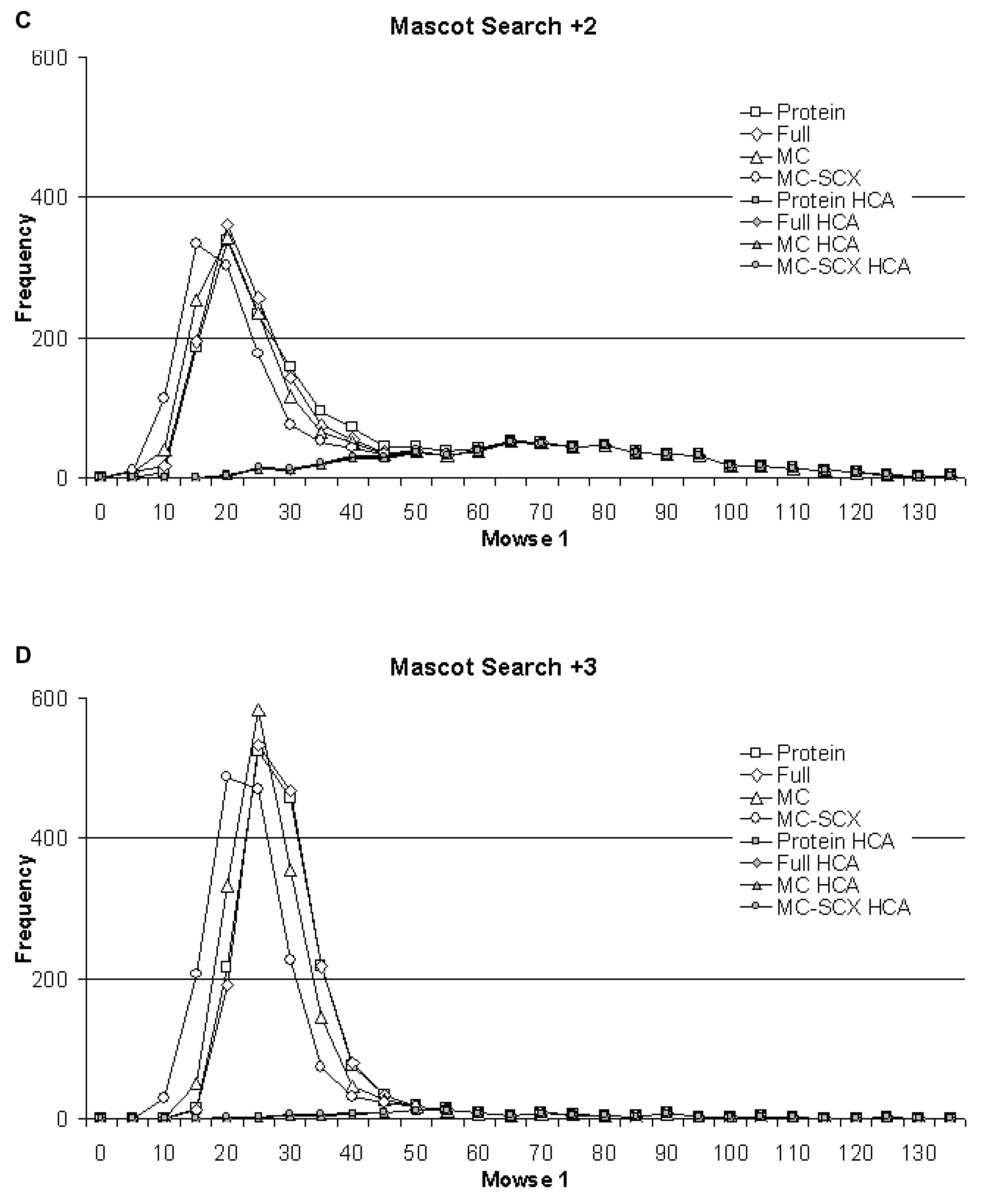


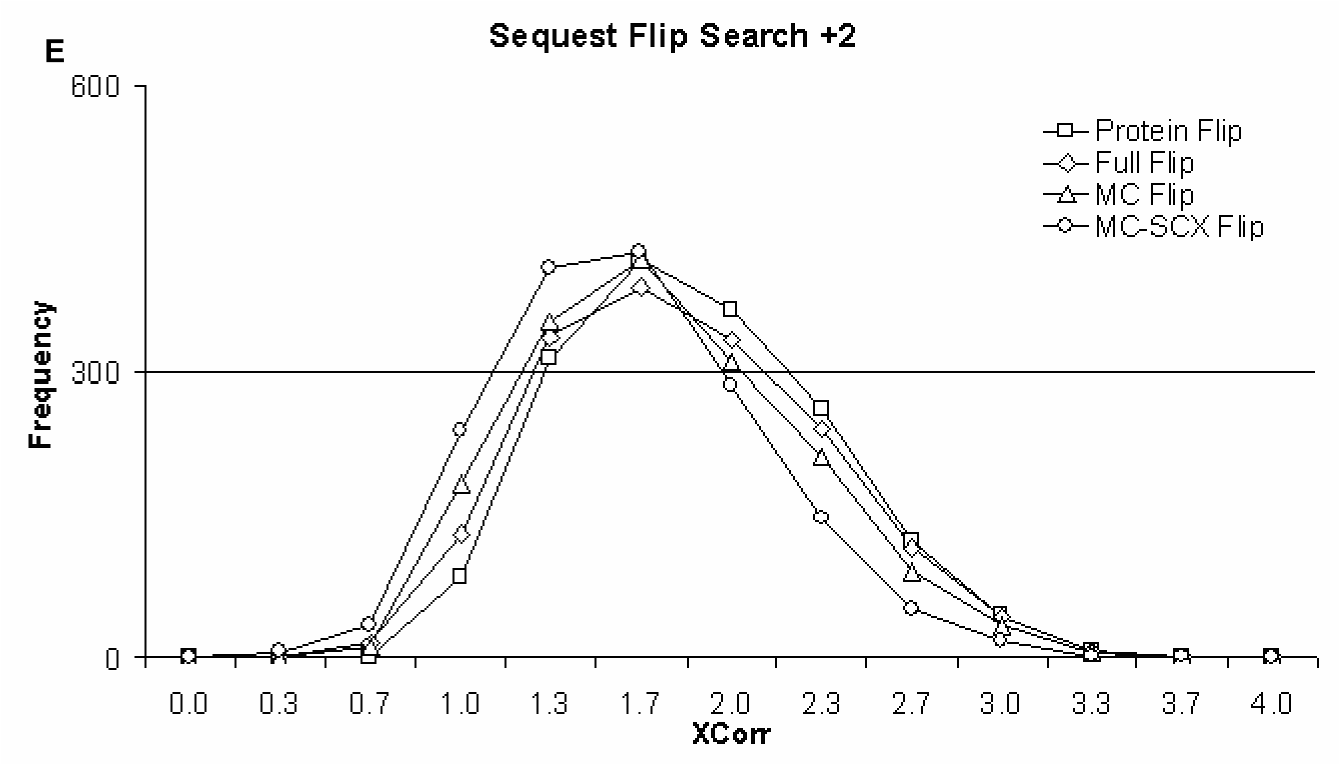
Supplementary Figure 1C. Comparison of no enzyme results

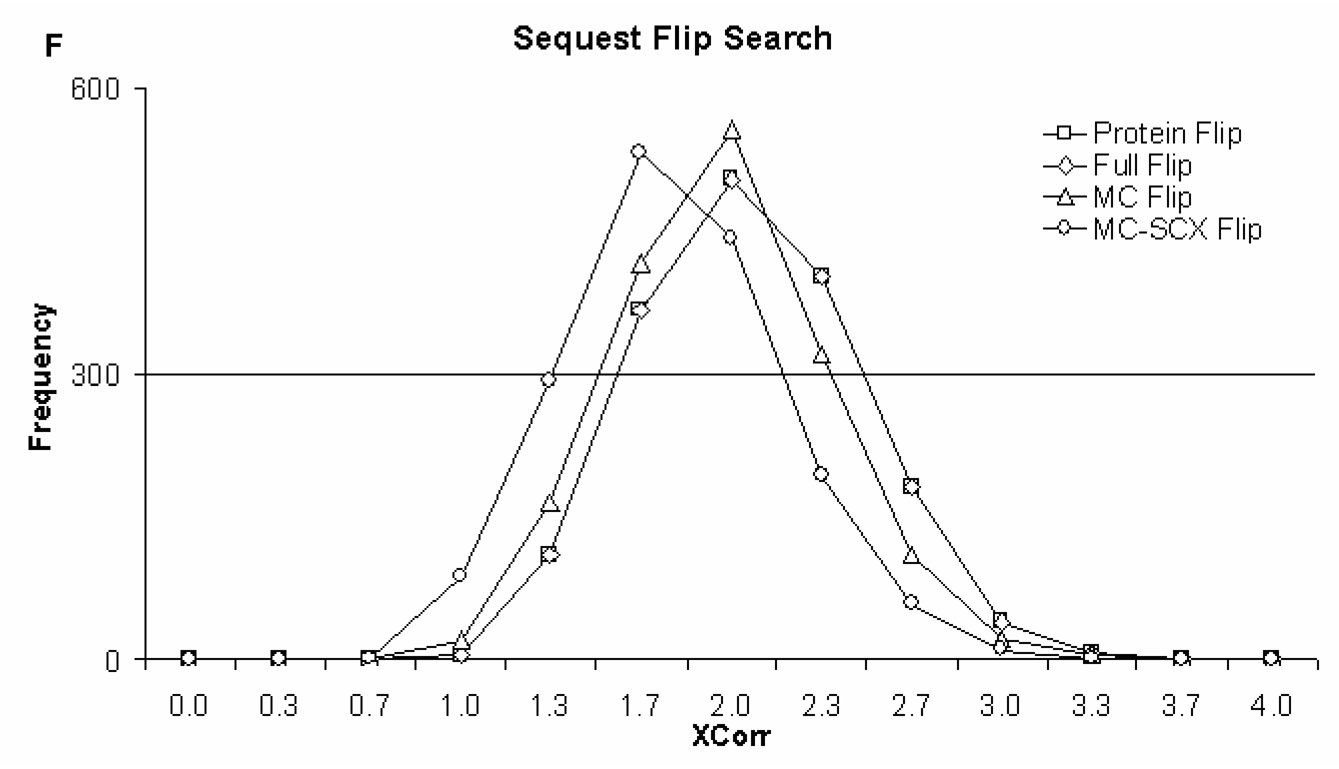


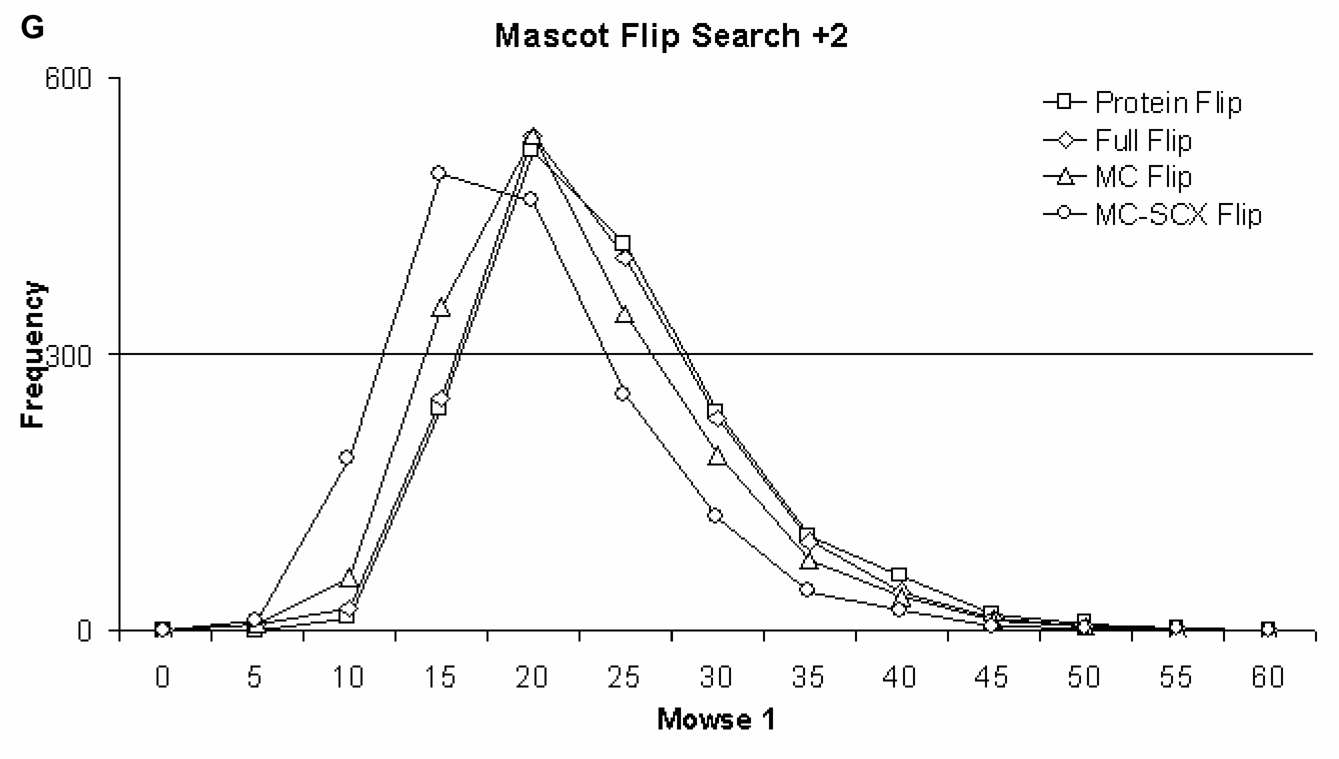
Supplementary Figure 2.

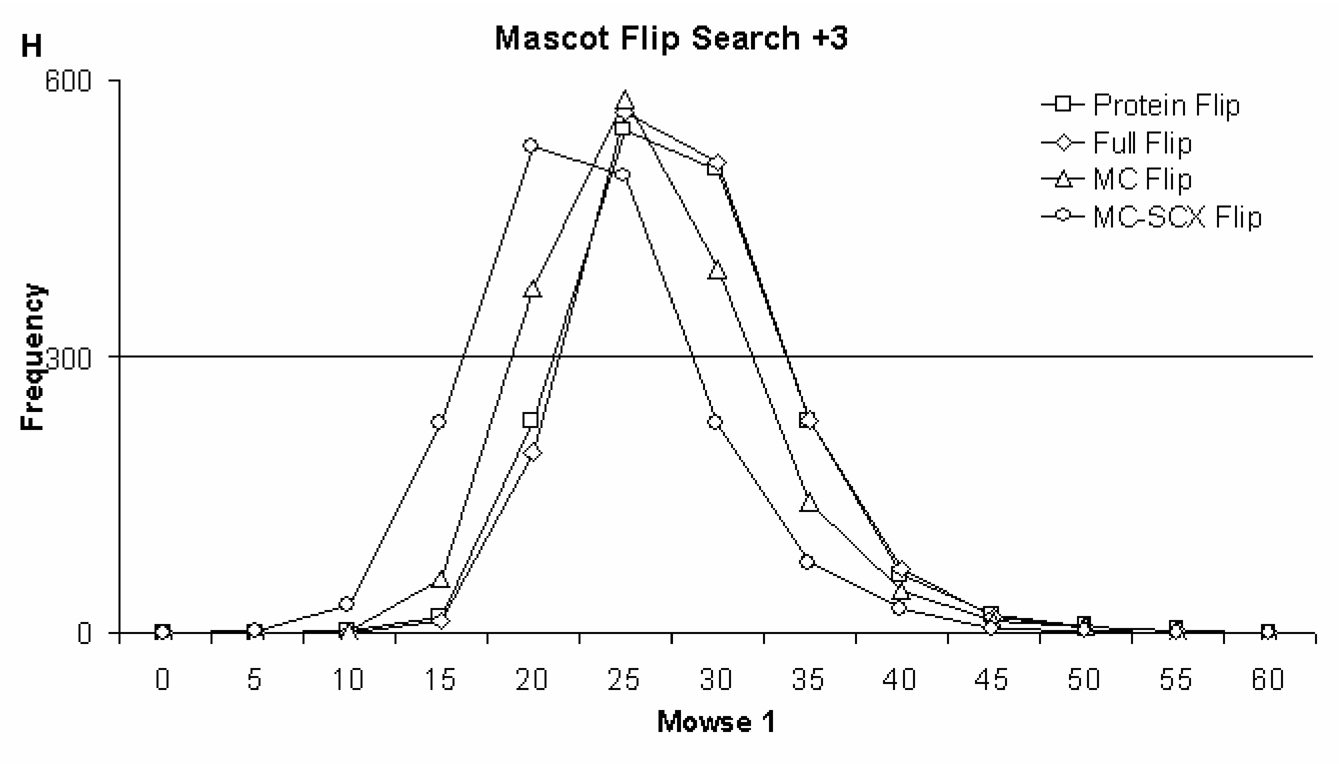



C
Mascot Search +2
Frequency
0
200
400
600
Mowse 1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Protein
Full
MC
MC-SCX
Protein HCA
Full HCA
MC HCA
MC-SCX HCA
D
Mascot Search +3
Frequency
0
200
400
600
Mowse 1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Protein
Full
MC
MC-SCX
Protein HCA
Full HCA
MC HCA
MC-SCX HCA


E
Sequest Flip Search +2
600
300
0
Frequency
0.0
0.3
0.7
1.0
1.3
1.7
2.0
2.3
2.7
3.0
3.3
3.7
4.0
XCorr
Protein Flip
Full Flip
MC Flip
MC-SCX Flip


F
Sequest Flip Search
600
300
0
Frequency
0.0
0.3
0.7
1.0
1.3
1.7
2.0
2.3
2.7
3.0
3.3
3.7
4.0
XCorr
Protein Flip
Full Flip
MC Flip
MC-SCX Flip


G
Mascot Flip Search +2
600
300
0
Frequency
Protein Flip
Full Flip
MC Flip
MC-SCX Flip
0
5
10
15
20
25
30
35
40
45
50
55
60
Mowse 1


H
Mascot Flip Search +3
600
300
0
Frequency
Protein Flip
Full Flip
MC Flip
MC-SCX Flip
0
5
10
15
20
25
30
35
40
45
50
55
60
Mowse 1