Supplementary figures, table and text

Improved Peptide Identification for Proteomic Analysis Based on Comprehensive Characterization of Electron Transfer Dissociation Spectra

Rui-Xiang Sun[1*], Meng-Qiu Dong[2*], Chun-Qing Song[2], Hao Chi[1], Bing Yang[2], Li-Yun Xiu[1], Li Tao[2], Zhi-Yi Jing[2], Chao Liu[1], Le-Heng Wang[1], Yan Fu[1], and Si-Min He[1]

[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[2] National Institute of Biological Sciences, Beijing 102206, China

*To whom correspondence should be addressed.

Email: rxsun@ict.ac.cn, or dongmengqiu@nibs.ac.cn

Running Title: ETD Characterization Improves Peptide Identification

Supplementary Figure S1	Distribution of the pFind scores and mass deviation for 23,649 ETD MS2 spectra with high-confidence identifications
Supplementary Figure S2	An ETD spectrum showing the typical histidine side chain loss
Supplementary Figure S3	ETD spectra of double, triply or quadruply charged peptides showing the different HR patterns
Supplementary Figure S4	The histograms of mass deviations for z, z+1, c-1, and c ions from the YEAST-ETcaD data set
Supplementary Figure S5	The histograms of mass deviations for y and b ions from the YEAST-B2R1 CID data set
Supplementary Figure S6	Notable side chain loss from z ions containing carbamidomethyl cysteines at their N-termini
Supplementary Figure S7	An example of a pair of ETD and CID MS2 spectra
Supplementary Figure S8	The ROC curves for ETD preprocessing and HR of doubly charged peptides identified by pFind
Supplementary Figure S9	Preprocessing by pRazor also improves the performance of Mascot on ETD data analysis
Supplementary Figure S10	Considerations of HR ions, in addition to the regular c and z ions, affect database search, in a precursor charge state dependent manner
Supplementary Figure S11	The histograms of mass deviations of z, z+1, c-1, and c ions from high-resolution and high-accuracy ETD and ECD data sets
Supplementary Table S1	Five AGC target settings and resulting MS2 spectra
Supplementary Text	Experimental optimization on AGC targets using a standard 8-protein mixture

Figure S1. Distribution of the pFind scores and mass deviation for 23,649 ETD MS2 spectra with high-confidence identifications.

These 23,649 ETD MS2 spectra correspond to the consensus peptides in Figure 1. The horizontal axis is the difference of the measured peptide mass from the calculated peptide mass. The pFind score on the vertical axis is -10*$\log_{10}$(E-value). Precursor mass deviations of these identified spectra are clustered tightly around zero (monoisotopic peaks) or 1 Da (isotopic peaks) as shown by the inset histograms.

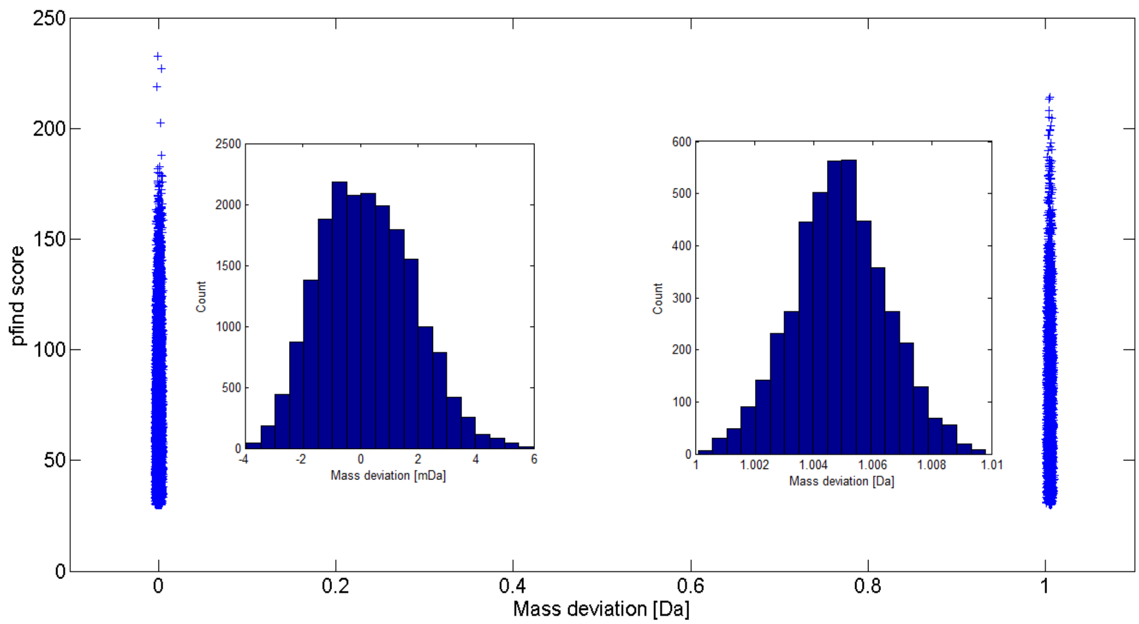


Figure S2. An ETD spectrum showing the typical histidine side chain loss. The spectrum is labeled and displayed using the software pLabel 2.0 (http://pfind.ict.ac.cn).

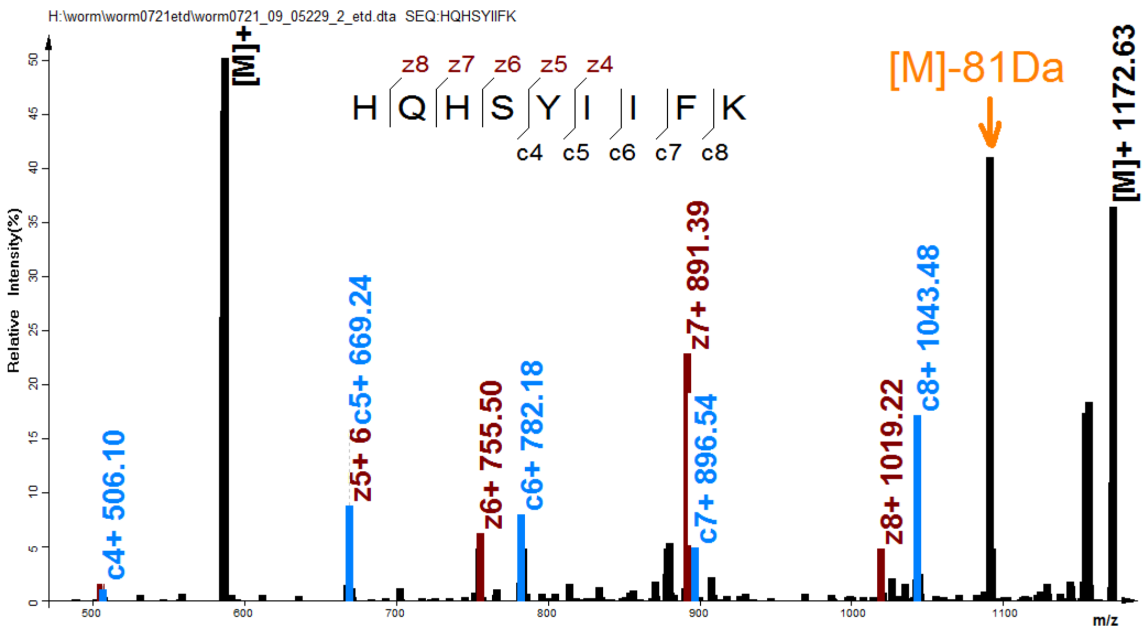
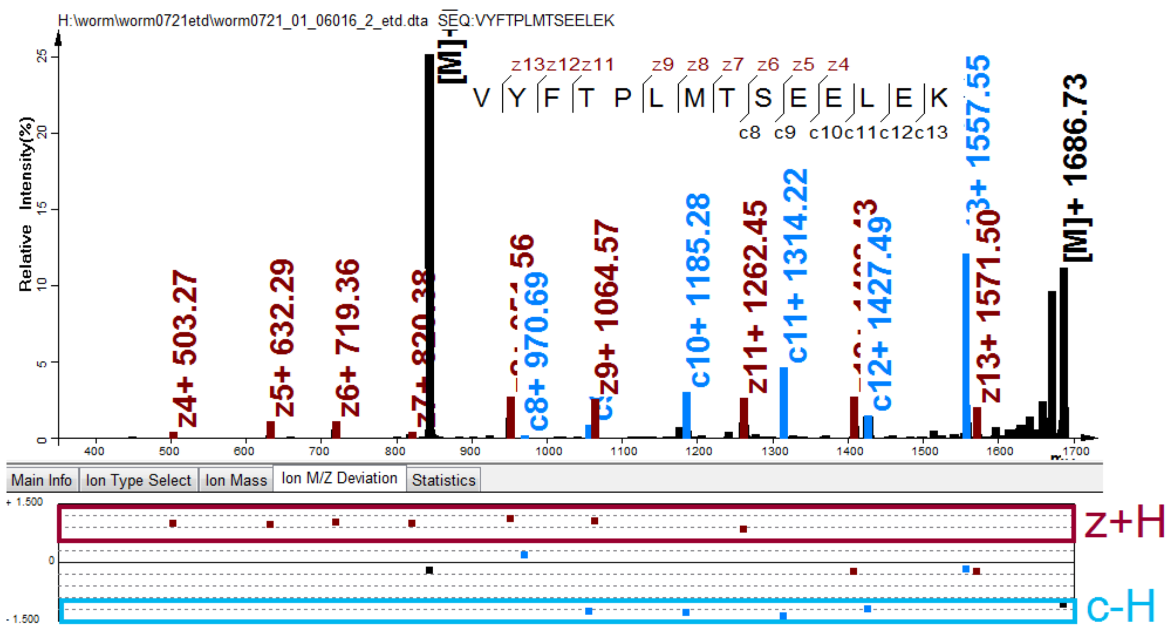
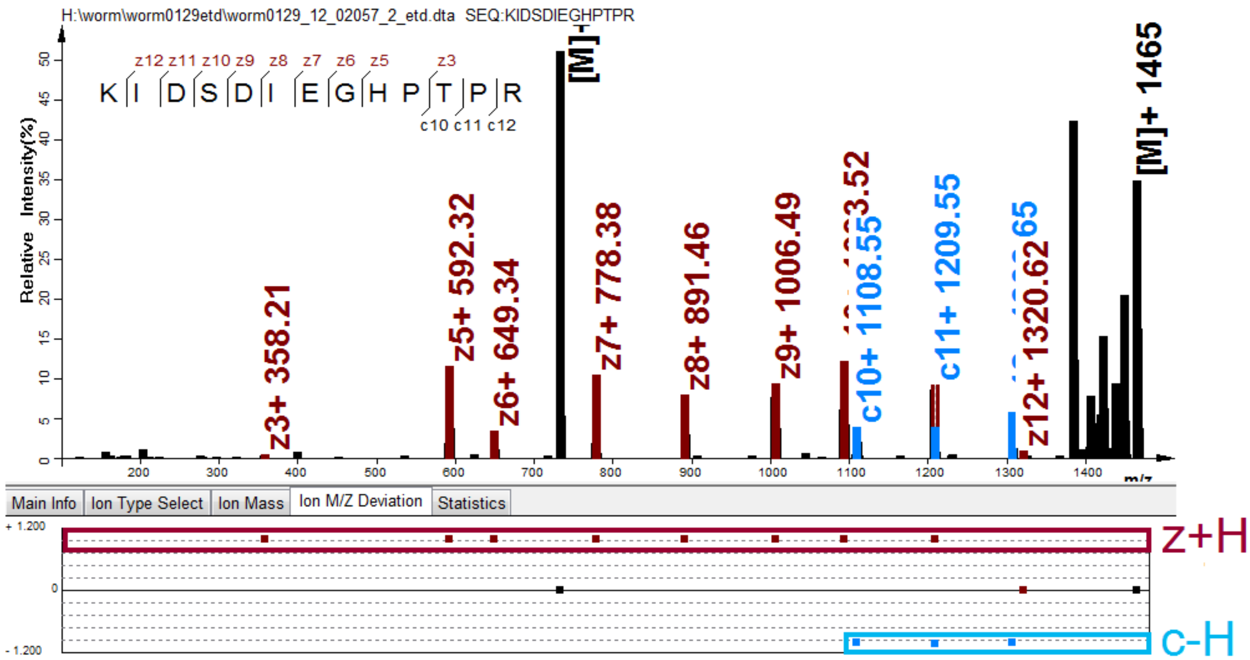


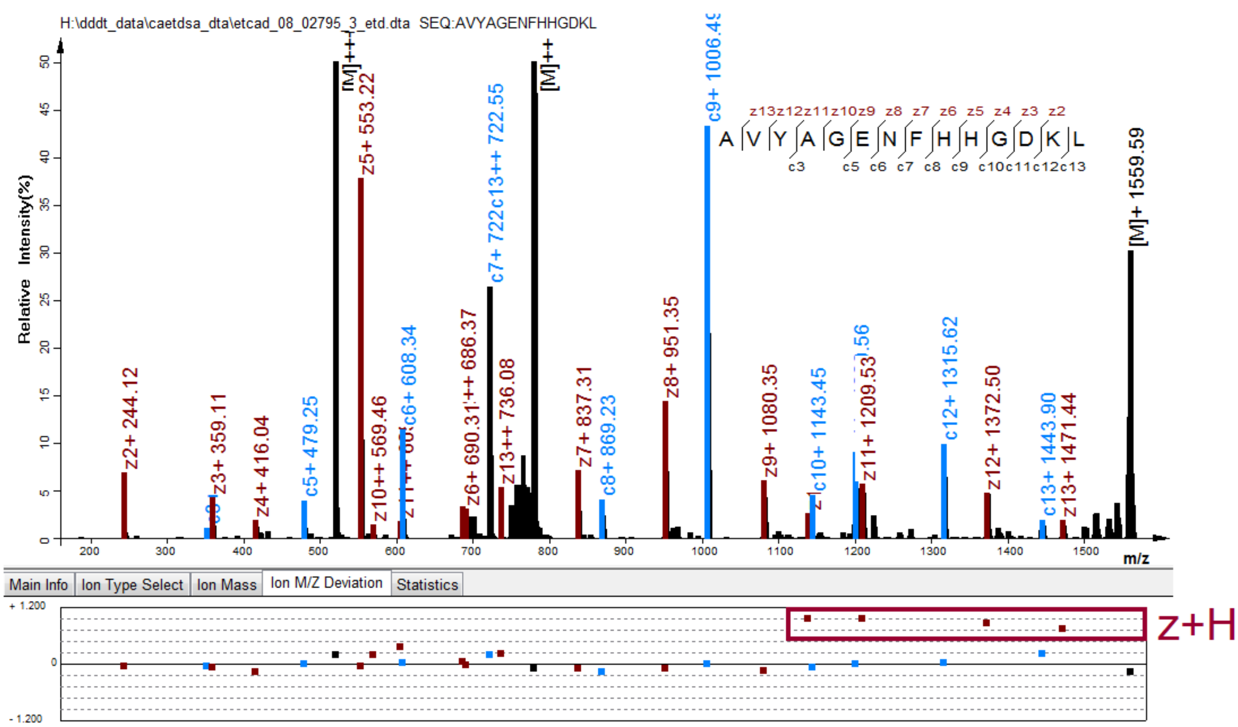
Figure S3. ETD spectra of double, triply or quadruply charged peptides showing the different HR patterns. (a) A normal-resolution ETD spectrum of doubly charged peptide VYFTPLMTSEELEK showing z+H and c-H ions. (b) A high-resolution ETD spectrum of doubly charged peptide KIDSDIEGHPTPR showing z+H and c-H ions. (c) A normal-resolution ETD spectrum of triply charged peptide AVYAGENFHHGDKL showing z+H ion and the lack of c-H ions. (d) A normal-resolution ETD spectrum of quadruply charged peptide RLEEDNDEEQSHKK showing z+H ion and the lack of c-H ions, similar to what is observed for +3 peptides. The rectangular window below each spectrum shows the mass deviation of each annotated peak.



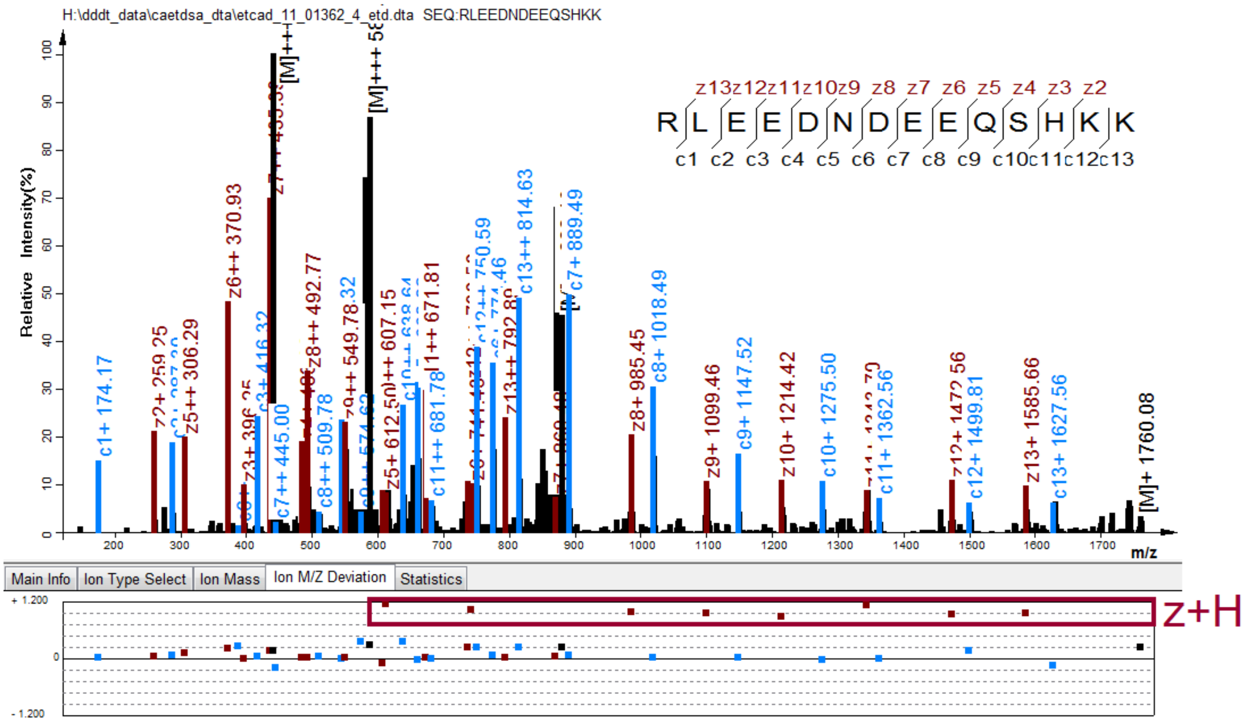
(a) A normal-resolution +2 ETD spectrum for peptide VYFTPLMTSEELEK



(b) A high-resolution +2 ETD spectrum for peptide KIDSDIEGHPTPR

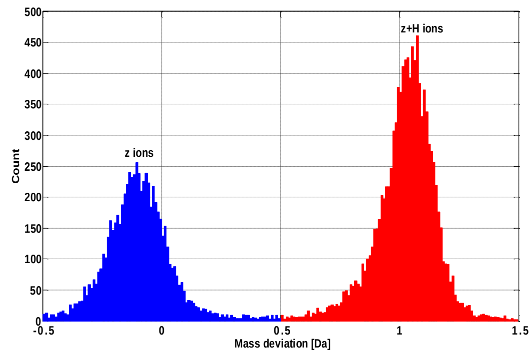


(c) A normal-resolution +3 ETD spectrum for peptide AVYAGENFHHGDKL

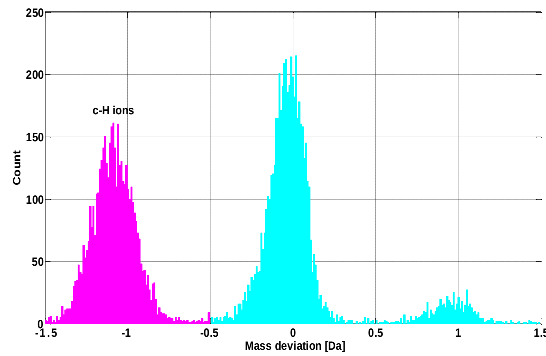


(d) A normal-resolution +4 ETD spectrum for peptide RLEEDNDEEQSHKK

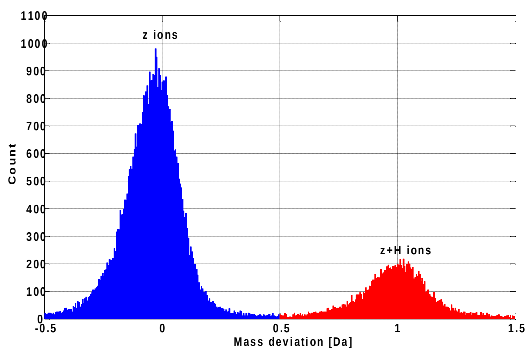
Figure S4. The histograms of mass deviations for z, z+1, c-1, and c ions from the YEAST-ETcaD data set. z and z+1 (a) or c-1 and c (b) in spectra of doubly charged peptides; z and z+1 (c) or c (d) in spectra of triply charged peptides; z and z+1 (e) or c (f) in spectra of quadruply charged peptides; z and z+1 (g) or c (h) in spectra of quintuply charged peptides.



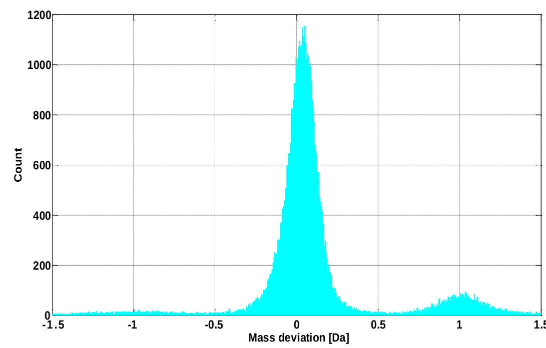
(a) z and z+1 ions (+2 peptides)



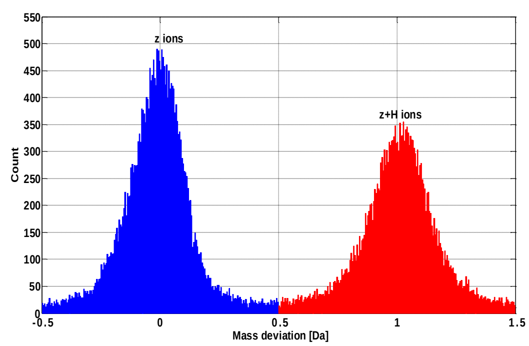
(b) c-1 and c ions (+2 peptides)



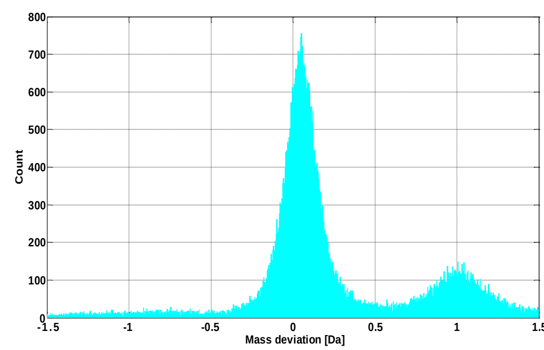
(c) z and z+1 ions (+3 peptides)



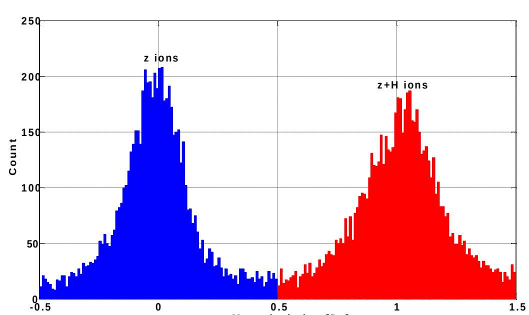
(d) c ions (+3 peptides)



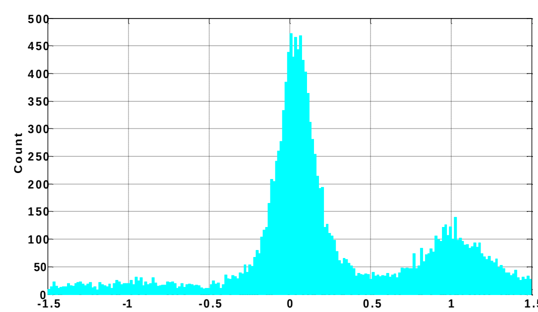
(e) z and z+1 ions (+4 peptides)



(f) c ions (+4 peptides)

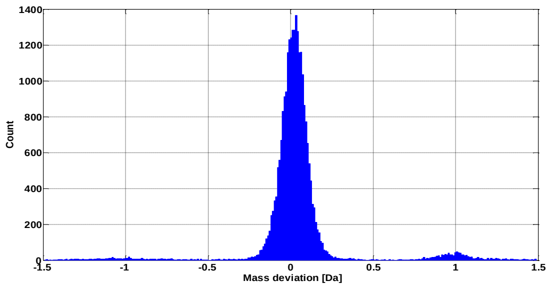


(g) z and z+1 ions (+5 peptides)

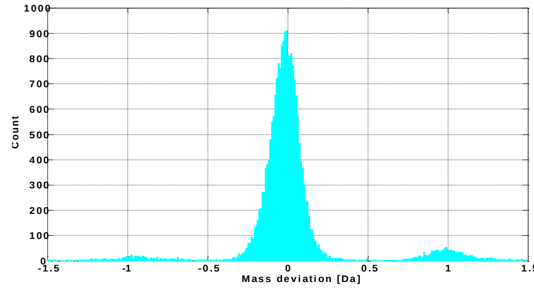


(h) c ions (+5 peptides)

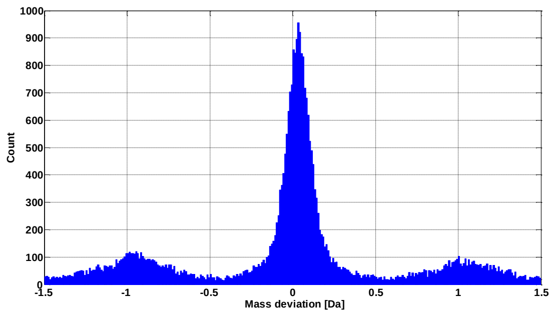
Figure S5. The histograms of mass deviations for y and b ions from the YEAST-B2R1 CID data set



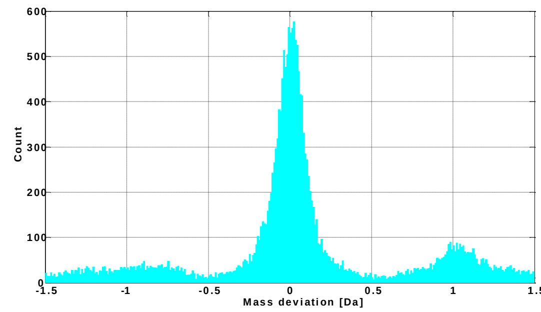
(a) y ions (+2 peptides)



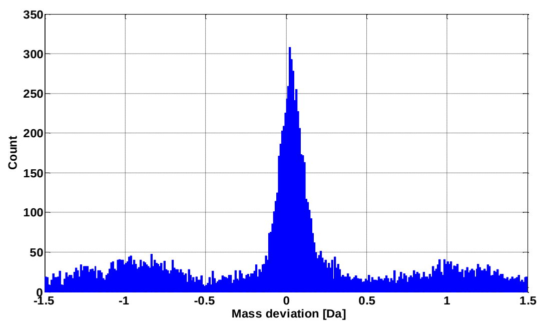
(b) b ions (+2 peptides)



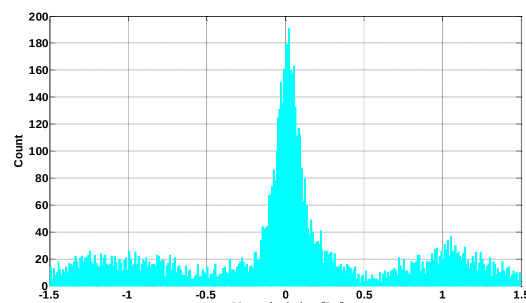
(c) y ions (+3 peptides)



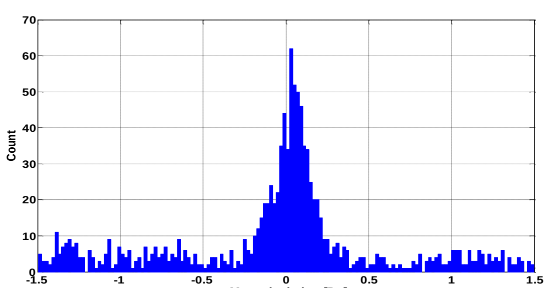
(d) b ions (+3 peptides)



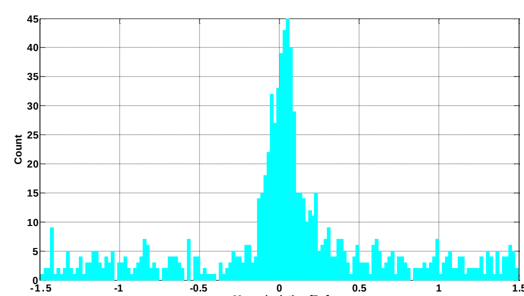
(e) y ions (+4 peptides)



(f) b ions (+4 peptides)

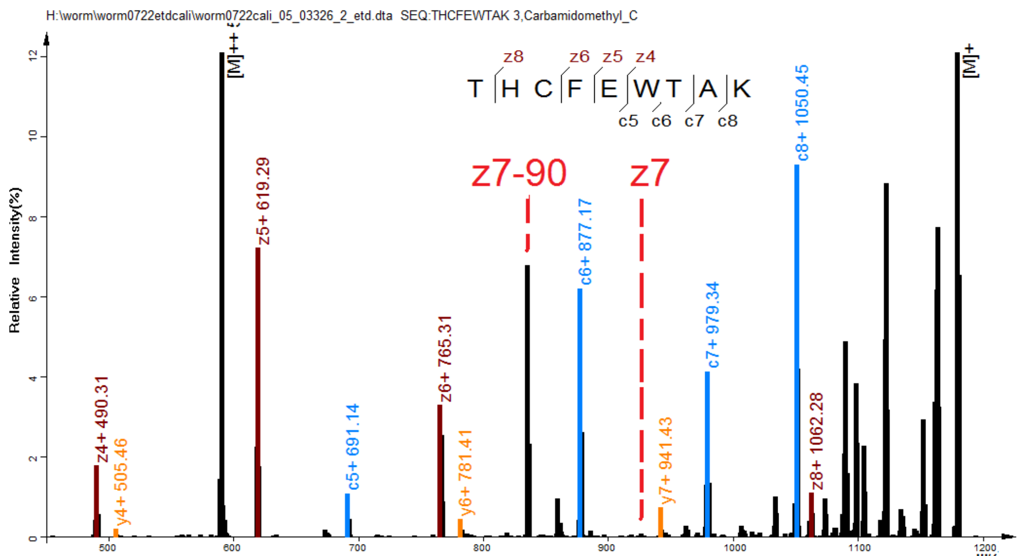


(g) y ions (+5 peptides)

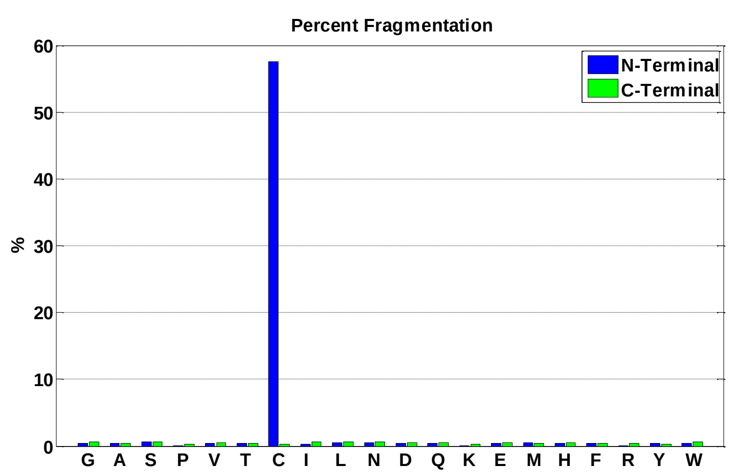


(h) b ions (+5 peptides)

Figure S6. Notable Side chain loss from z ions containing carbamidomethyl cysteines at their N-termini. (a) The dotted line in the ETD spectrum of THCFEWTAK points to the z7 ion with a prominent side chain loss of ~90 Da (z7-90), whereas the peak of intact z7 was not observed. (b) Occurrence frequencies of z-90 Da peaks that are generated either N- or C-terminal to the indicated amino acid. The statistical results were from the SwedECD data set.



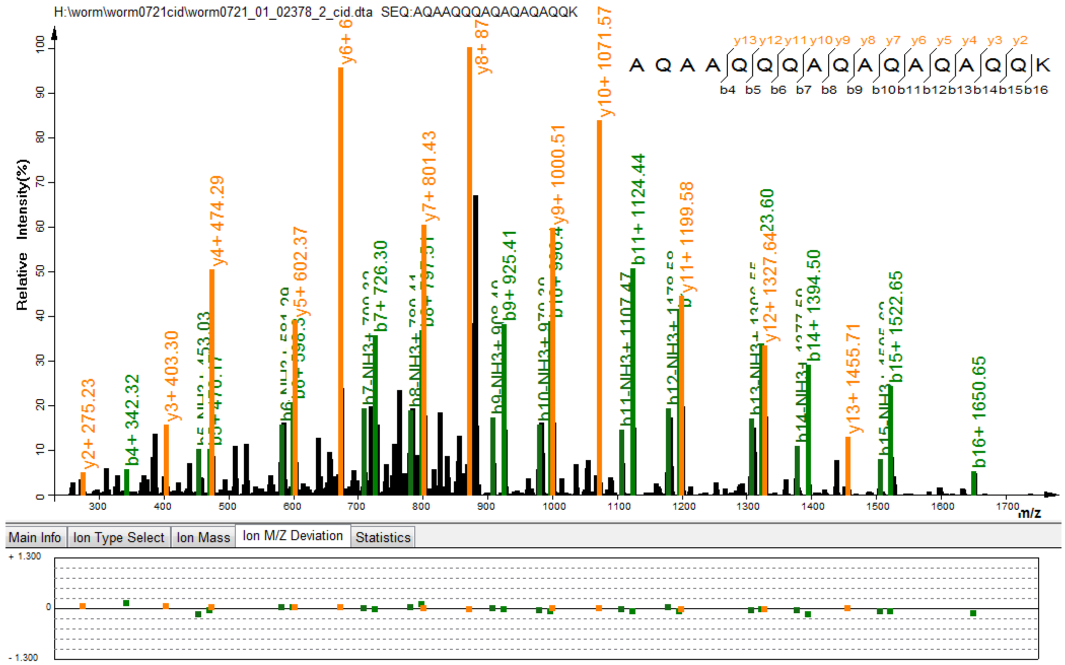
(a) An ETD spectrum with the dot lined peak corresponding to the -90Da mass loss of z_7



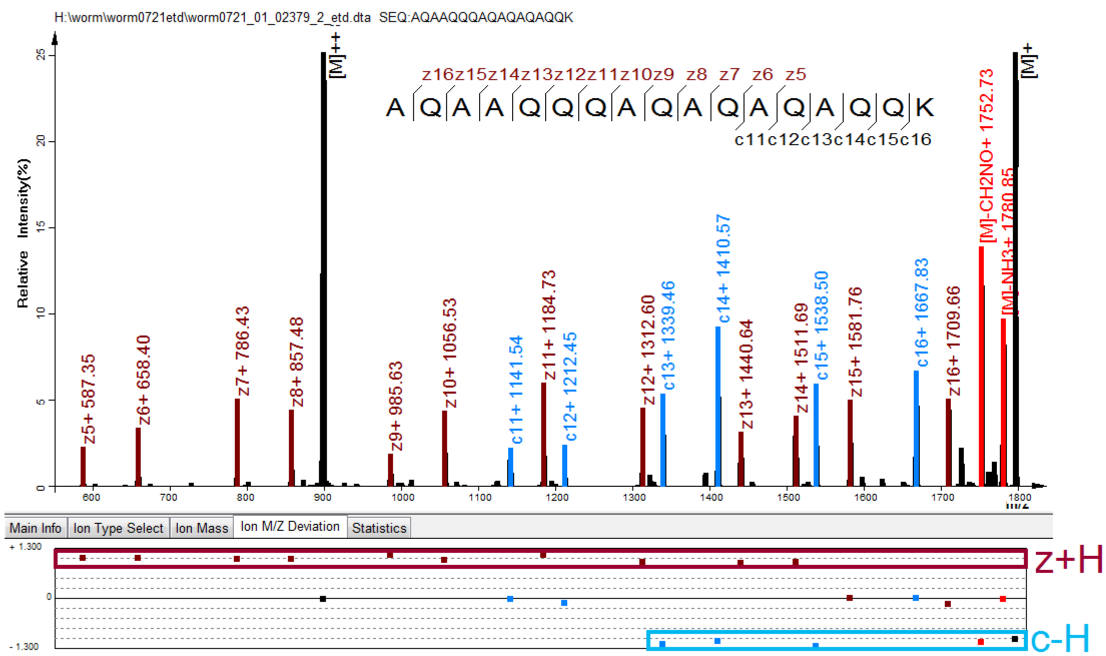
(b) Occurrence frequencies of z-90 Da peaks for 20 AAs

Figure S7. An example of a pair of ETD and CID MS2 spectra.

The CID spectrum (a) and ETD spectrum (b) of a doubly charged peptide AQAAQQQAQAQAQAQQK.



(a) CID



(b) ETD

Figure S8. The ROC curves for ETD preprocessing and HR of doubly charged peptides identified by pFind. ETD data were preprocessed (w/ preprocessing) or not (w/o preprocessing) using pRazor to remove high-intensity un-reacted precursors, CR precursors and related neutral loss species. "c&z only" means that only regular c and z ions were considered in spectral matching. "c-H", z+H" or "c-H, z+H" indicate that in addition to regular c and z ions, c-H, z+H, or both c-H and z+H ions were considered in spectral matching, respectively.

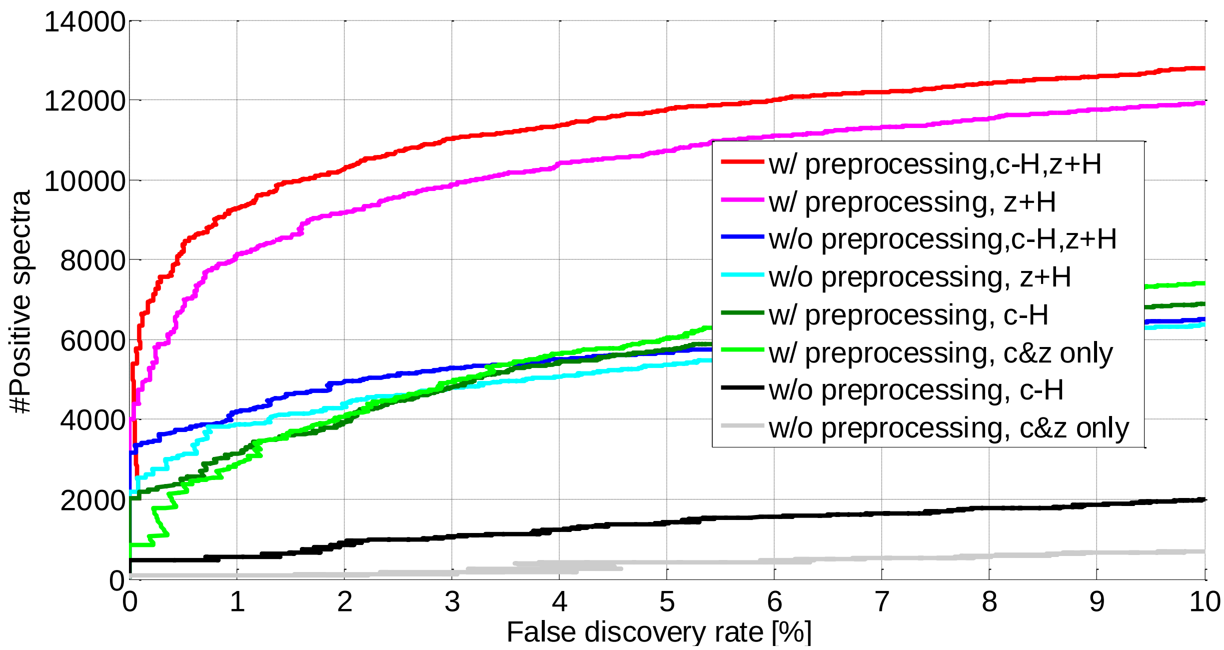
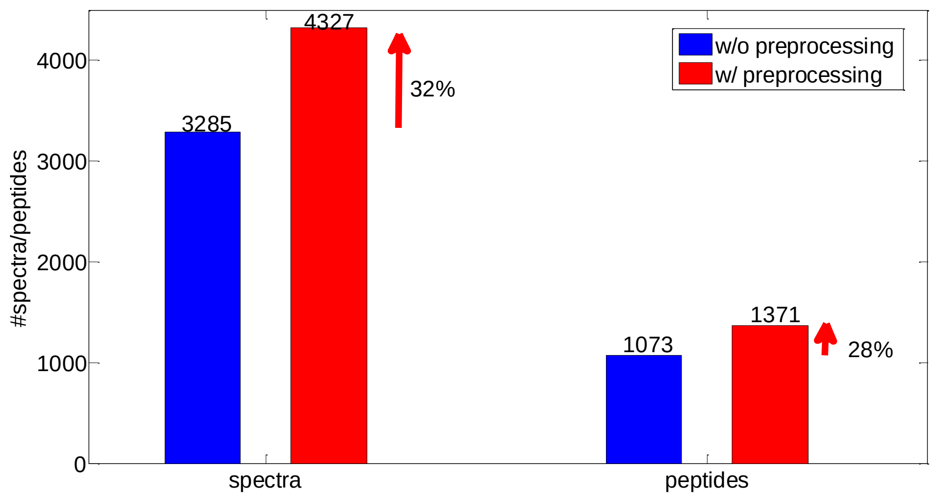
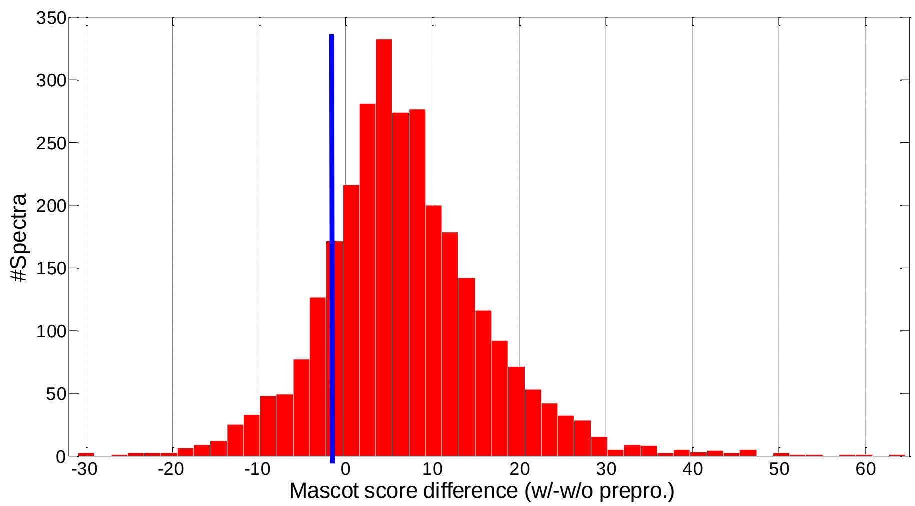


Figure S9. Preprocessing by pRazor also improves the performance of Mascot on ETD data analysis. (a) With pRazor preprocessing, Mascot identified 32% more ETD spectra, which corresponds to an increase of 28% more peptides, from WORM-A data set. (b) For the majority of the identified spectra, pRazor preprocessing improved their Mascot scores. Distribution of Mascot score difference (w/ preprocessed minus w/o preprocessed). Spectra with improved Mascots scores are located on the right side of the blue line.



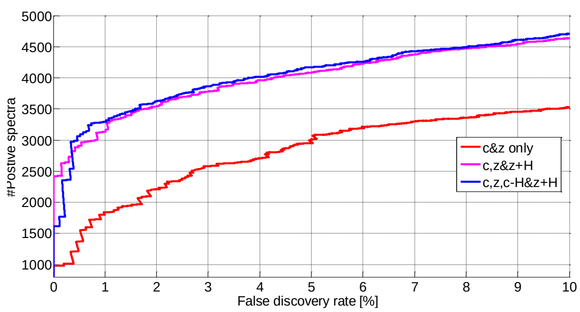
(a)



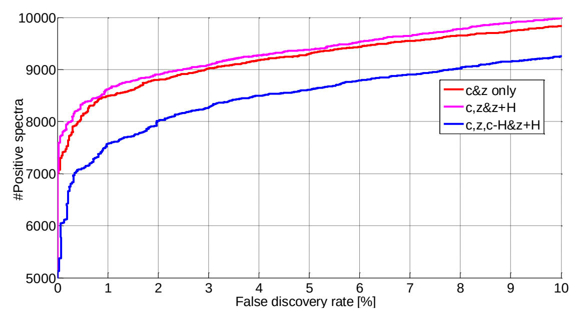
(b)

Figure S10. Considerations of HR ions, in addition to the regular c and z ions, affect database search, in a precursor charge state dependent manner.

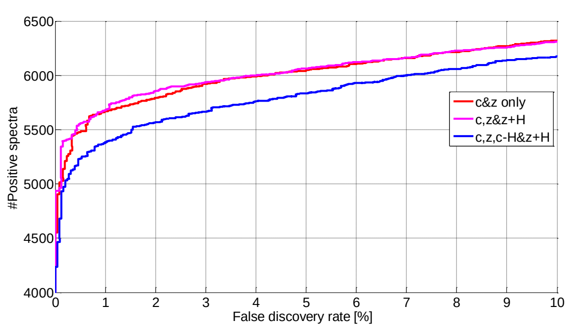
The YEAST-ETcaD data were analyzed using pFind. In the ROC curve plots, “c&z only” means that only regular c and z ions were considered in spectral matching; “c,z&z+H” means that in addition to regular c and z ions, z+H ions were considered; “c,z,c-H&z+H” indicates that in addition to regular c and z ions, both c-H and z+H ions were considered in spectral matching. For +2 peptides the best performance was achieved using “c,z,c-H&z+H”, while for +3, +4, or +5 peptides, it is most desirable to use “c,z&z+H”.



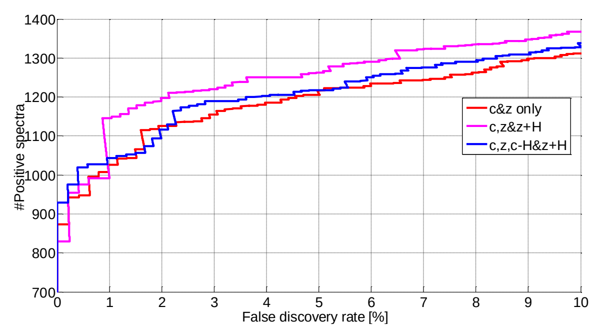
(a) +2 precursors



(b) +3 precursors

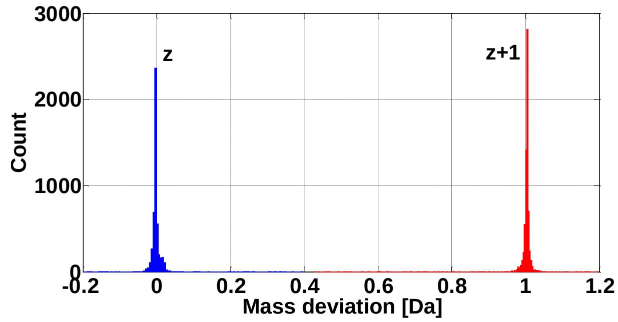


(c) +4 precursors

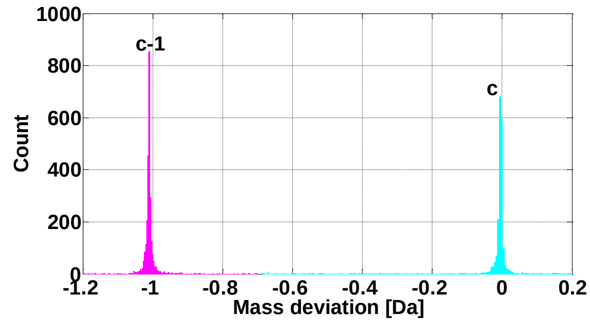


(d) +5 precursors

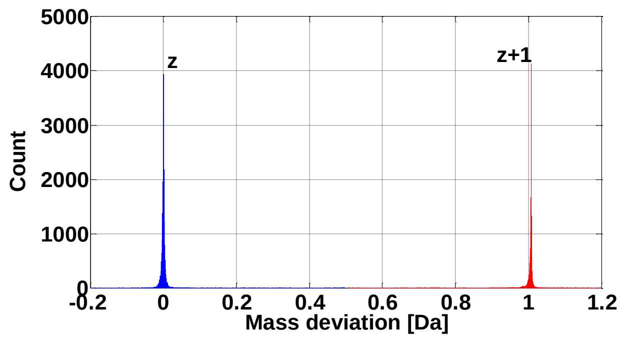
Figure S11. The histograms of mass deviations of z, z+1, c-1, and c ions from high-resolution and high-accuracy ETD and ECD data sets. (a) z and z+1 ions (WORM-H); (b) c-1 and c ions (WORM-H); (c) z and z+1 ions (SwedECD); (d) c-1 and c ions (SwedECD)



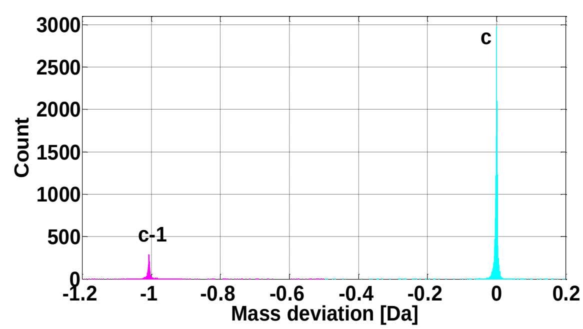
(a) z and z+1 ions (WORM-H ETD)



(b) c-1 and c ions (WORM-H ETD)



(c) z and z+1 ions (SwedECD)



(d) c-1 and c ions (SwedECD)

Supplementary Table S1. Five AGC target settings and resulting MS2 spectra

No.	AGC(LTQ)	AGC(OT)	AGC(reagent)	#Spec acquired	#Spec ID CID-IT	#Spec ID ETD-IT	#Spec ID ETD-OT
Set1	1e4	1e5	1e5	2085x3	285	237	203
Set2	***2e4***	***2e5***	***1e5***	***2184x3***	***421***	***334***	***279***
Set3	4e4	4e5	1e5	2028x3	274	211	209
Set4	2e4	2e5	2e5	1969x3	330	275	226
Set5	2e4	2e5	4e5	1952x3	346	276	218

Supplementary Text

Experimental optimization on AGC targets using a standard 8-protein mixture

The primary goal of this experiment is to optimize AGC targets to acquire MS2 spectra that would result in as many high-confidence identifications as possible. Five settings were tested (see supplementary Table S1). The sample is a tryptic digest of eight standard proteins which are Rabbit skeletal muscle Myosin, Beta-Galactosidase of *E.coli*, Rabbit muscle Phosphorylase b, Bovine Serum albumin, Chicken Ovalbumin, Bovine Carbonic anhydrase, Soybean Trypsin inhibitor, and Chicken Lysozyme. The experiment was carried out on an LTQ-Orbitrap XL with ETD (Thermo Scientific). Different combinations of the automatic gain control (AGC) for LTQ MS2, Orbitrap (OT) MS1 and ETD reagent ion (Fluoranthene) are shown in Table S1. Each of the three most intense precursors is sequentially dissociated by CID, ETD, and ETD. The first CID as well as the first ETD spectrum is of normal resolution generated in LTQ (CID-IT and ETD-IT, respectively). The second ETD spectrum is of high resolution produced in Orbitrap (ETD-OT). All of the MS2 spectra were identified by the search engine pFind 2.1 against the eight protein sequences with the yeast protein sequences appended to them. As the AGC targets of Set2 gave the best performance (bold and italic in Table S1), we used Set2 in the triplicate worm CID/ETD ‘double play’ experiments.