Supplementary Information

to

PTMiner: localization and quality control of protein modifications detected in an open search and its application to comprehensive PTM characterization in human proteome

Zhiwu An[1,4,*], Linhui Zhai[2,*], Wantao Ying[3], Xiaohong Qian[3], Fuzhou Gong[1,4,#], Minjia Tan[2,#], Yan Fu[1,4,#]

[1]*National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[2]*State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

[3]*State key Laboratory of Proteomics, National Center for Protein Sciences Beijing, Beijing Proteome Research Center, National Engineering Research Center for Protein Drugs, Beijing 102206, China, Beijing Institute of Lifeomics, Beijing 100850, China*

[4]*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*

*These authors contributed equally to this work

#Correspondence should be addressed to:

Yan Fu (yfu@amss.ac.cn)

Minjia Tan (mjtan@simm.ac.cn)

Fuzhou Gong (fzgong@amt.ac.cn)

Supplementary Notes

Supplementary Note 1. If we use the global FDR approach to calculate the FDR of the whole identification results, the real FDR is almost equal to the estimated FDR (Supplementary Fig. 1A). However, the group FDRs of different modification groups at the same identification scores are different (Supplementary Fig. 1B-D). As illustrated in our previous paper [1], the group FDR can be calculated from the global FDR using the formula below,

$$FDR_k(x) = \frac{FDR(x)}{FDR(x) + \frac{\lambda_k(x)}{\gamma_k(x)}(1 - FDR(x)}$$

where symbol k denotes one type of modification, x indicates identification score threshold, $FDR_k(x)$ and $FDR(x)$ indicate the group FDR of PSMs carrying modification k and the global FDR for the PSMs owning identification scores greater than x, respectively, $\lambda_k(x)$ is the probability that a spectrum is identified as a k-modified peptide given that the identification is true and the score is greater than x, and $\gamma_k(x)$ is the probability that a spectrum is identified as a k-modified peptide given that the identification is false and the score is greater than x.

From this formula we know that FDRs of different modifications are different at the same score threshold because $\lambda_k(x)$ and $\gamma_k(x)$ are different for them. $\lambda_k(x)$ is directly related to the proportion of spectra that are produced from k-modified peptides in the protein sample, while $\gamma_k(x)$ is closely related to the proportion of candidate peptides with modification k in the search space. In an open search, the abundances of peptides with different mass shifts in the search space are similar but the abundances in the spectra can be dramatically different, resulting in the heterogeneity of FDRs for different mass shifts at the same score level.

Supplementary Note 2. In order to determine the initial values of Gaussian mixture models when clustering the mass shifts in each 1-Da bin, a discrete convolution method is introduced. The mass shifts in each bin are first grouped into small windows (the window size is set to one percentage of the precursor tolerance *T*), and then the discrete convolution algorithm will be carried out,

$$\mu(k) = \sum_{j=0}^{2n} v(k-j)\,\omega(j) \tag{1}$$

where $n = 100$, ω is a $2n+1$ dimensional weight vector, v and μ are the number vectors of mass shifts in each small window before and after the convolution operation, respectively, and $k = 1,2,\ldots,n/T$. We assume $v(k-j)$ is zero if the index $k-j$ is not positive.

The weight vector is set using a Gaussian kernel function,

$$\omega(j) = exp\left\{-\frac{(j-n)^2}{2\sigma^2}\right\} \tag{2}$$

where $\sigma = n/6$, and $j = 0,1,\ldots,2n$. In doing so, $\mu(k)$ can be regarded as a weighted average number of mass shifts in the mass interval of $[M(k)-T, M(k)+T]$, where $M(k)$ is the central mass of the *k-th* small window. Every mass center $M(k)$ has a corresponding mass variance $V(k)$, which is estimated as the variance of mass shifts falling into the interval of $[M(k)-T, M(k)+T]$. We check each $M(k)$ in descending order of $\mu(k)$. If its corresponding variance is less than 1.5 times of the system measure variance, which is estimated as the variance of mass shifts falling into the interval of [*-T, T*], then $M(k)$ will be regarded as one candidate modification mass. Note than these $M(k')$ within the mass interval of $[M(k)-T, M(k)+T]$ will not be considered any more.

In this way, we will get the maximal number (*K*) of potential modification types as well as the initial means (*Ms*) and variances (*Vs*) of the *K* clusters in each bin.

Supplementary Note 3. To generate the simulated spectra, we first used a Markov chain model, which was trained on the Swiss-Prot human protein sequences (downloaded on October 28, 2015), to generate the random protein sequences as described previously[2]. A total of 20,000 protein sequences were generated as the target database, and additional 20,000 protein sequences were generated to produce contaminated spectra. Then, the two sets of sequences were theoretically digested by trypsin with up to 2 allowed miss-cleavages, resulting in a theoretical peptide set. Finally, we randomly selected a part of peptide sequences from the peptide set to generate their simulated spectra using a peak-replacing method described below. (a) We used pFind to search a real data set (samples Adult_Adrenalgland_Gel_Elite_49 and Adult_Adrenalgland_Gel_Velos_2 from the human proteome map [3]) in the open search mode, and selected the unmodified PSMs to build up a PSM collection. (b) For a theoretical peptide sequence, we randomly picked out a PSM with the same peptide length and charge state in the PSM collection, and moved the matched peaks to the m/z positions of corresponding fragment ions of the theoretical peptide, retaining the same intensities and mass errors. The peptide precursor mass was also changed to the mass of the theoretical peptide with the same mass error. The unmatched peaks were retained.

In order to evaluate how realistic the simulated spectra were, we compared them to the experimentally acquired spectra. We selected one sample of the human proteome map (Adult_Bcells_Gel_Elite_76, we call it real data below). The identified target PSMs without filtering were used. Because the lengths of the theoretical peptides were between 6 and 30 and their charges were 2 or 3, those PSMs with peptide lengths larger than 30 or charges larger than 3 in the real data were removed. In addition, because we did not know exactly which amino acid one mass shift occurred on, we next removed the modified PSMs (mass shifts outside the interval of [-0.5, 0.5] Da) from both data sets. Supplementary Fig. 4 shows the similar distribution properties of the two kinds of data.

Supplementary Note 4. The modified-peptide spiked-in complex proteome data was produced as follows. The synthetic peptides were dissolved in buffer (2% acetonitrile in 0.1% formic acid) at a concentration of 10 pmol/μL per peptide. In total, the synthetic peptide mixture (50 pmol of each one) was mixed with 1μg trypsin digested HeLa or *E. coli* whole cell proteins, respectively. The peptide samples were analyzed by nano-liquid chromatography-tandem mass spectrometry. The peptides were separated and eluted from home-made C18 reverse-phase analytical column (75um i.d. X 15cm length) with a 120min linear gradient. The particle size of the C18 resin was 3μm (100 Å，Dikma Technologies, Lake Forest, CA). The eluted peptides were ionized and analyzed by Q Exactive mass spectrometer (Thermo Fisher Scientific, Waltham, MA). For the MS1 analysis, the full scan range was set 350-1300 m/z, the resolution was set at 70000 (m/z 200), and the automatic gain control (AGC) was set $1x10^6$. The 16 most intense ions were selected to be fragmented via HCD with the normalized collision energy (NCE) of 28%. The fragment ions were detected by Orbitrap at the resolution of 17500 (200 m/z). The AGC was set $1x10^6$, and the dynamic exclusion was set 40 s. Finally, two raw data files including a total of 88,146 tandem mass spectra were obtained.

Supplementary Figure

Supplementary Fig. 1. The FDRs of unmodified and differently modified peptides are likely to be different at the same score threshold.

Supplementary Fig. 2. Example of $\gamma_k(x)$ in transfer FDR that was well fitted by linear regression.

Supplementary Fig. 3. Example of relative intensities of matched peaks of unmodified PSMs, which were fitted using lognormal distribution.

Supplementary Fig. 4. The evaluation of realness of the simulated data.

Supplementary Fig. 5. Comparison results of the original Ascore and our extended Ascore on phosphorylated PSMs.

Supplementary Fig. 6. The real FLRs are plotted against the PTMiner posterior probability and Ascore for the open search results of the simulated data set.

Supplementary Fig. 7. Examples of identified peptides that do not belong to the chemically synthesized modified peptides in the ProteomeTools project.

Supplementary Fig. 8. The FDRs of different modifications are different at the same score threshold for the data set of draft map of human proteome.

Supplementary Fig. 9. The two examples of post-translational modifications (A is for Oxidation [P] and B is for Di-oxidation [M]) identified and localized with high confidence from the draft map of human proteome.

Supplementary Fig. 10. Two examples of MS/MS spectra of peptides with deamidation at Arg and succinylation at Lys.

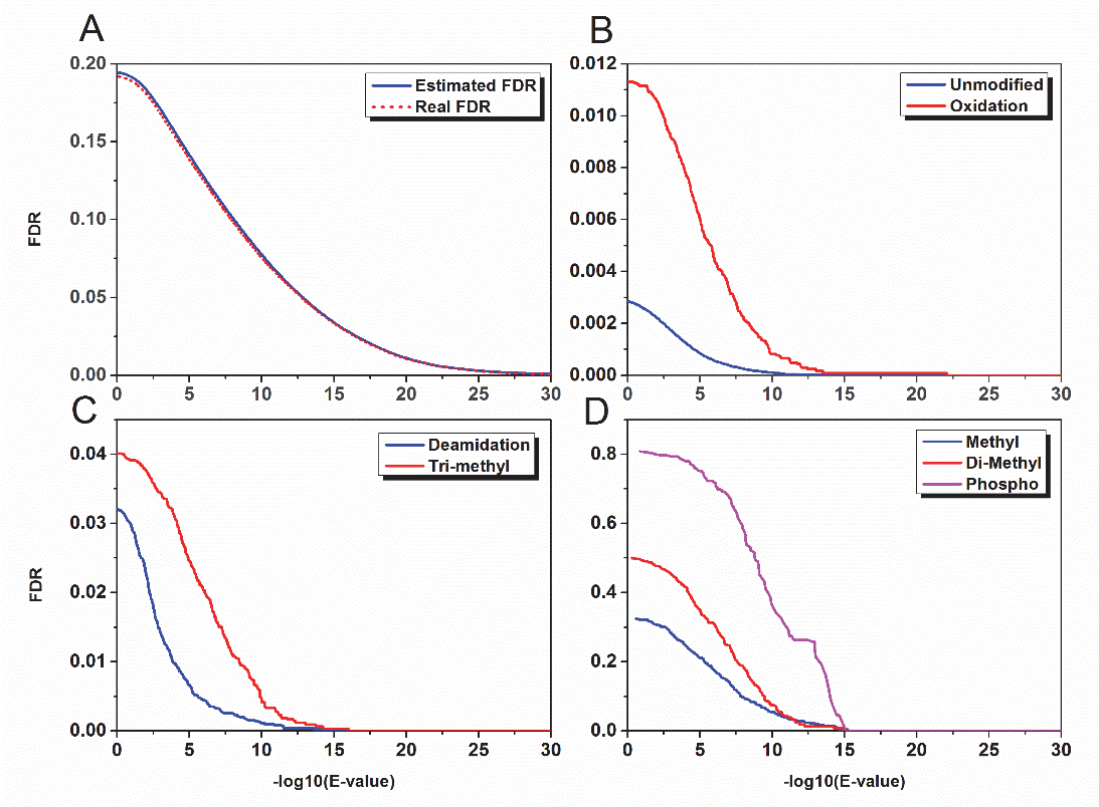
Supplementary Fig. 11. One PSM example for benzoylation modification.

Supplementary Fig. 12. Two examples of MS/MS spectra of peptides with single amino acid variations (SAVs) annotated in the UniProt database.

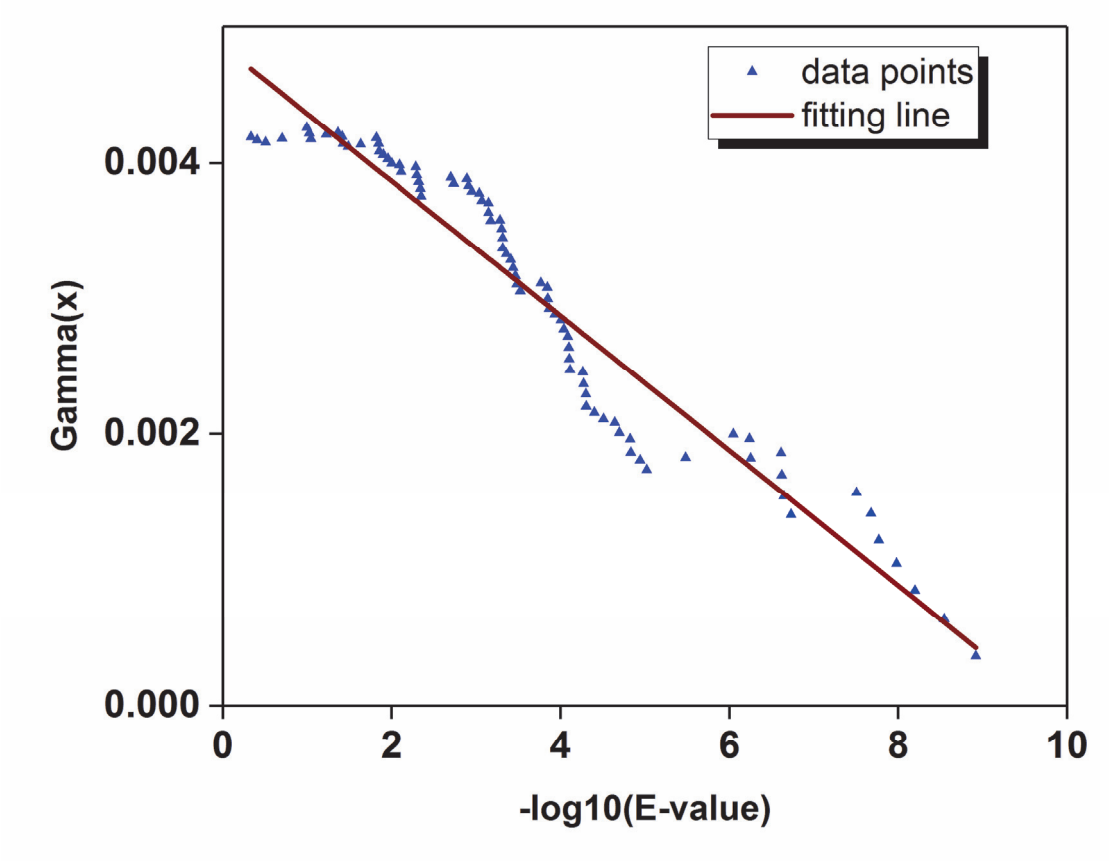
Supplementary Fig. 13. Some fully annotated modifications show strong differences between sample preparation methods (SDS-PAGE and bRP).

Supplementary Fig. 14. Twenty representative PSMs with mass shift of 12.000000 Da localized to peptide N-terminal.

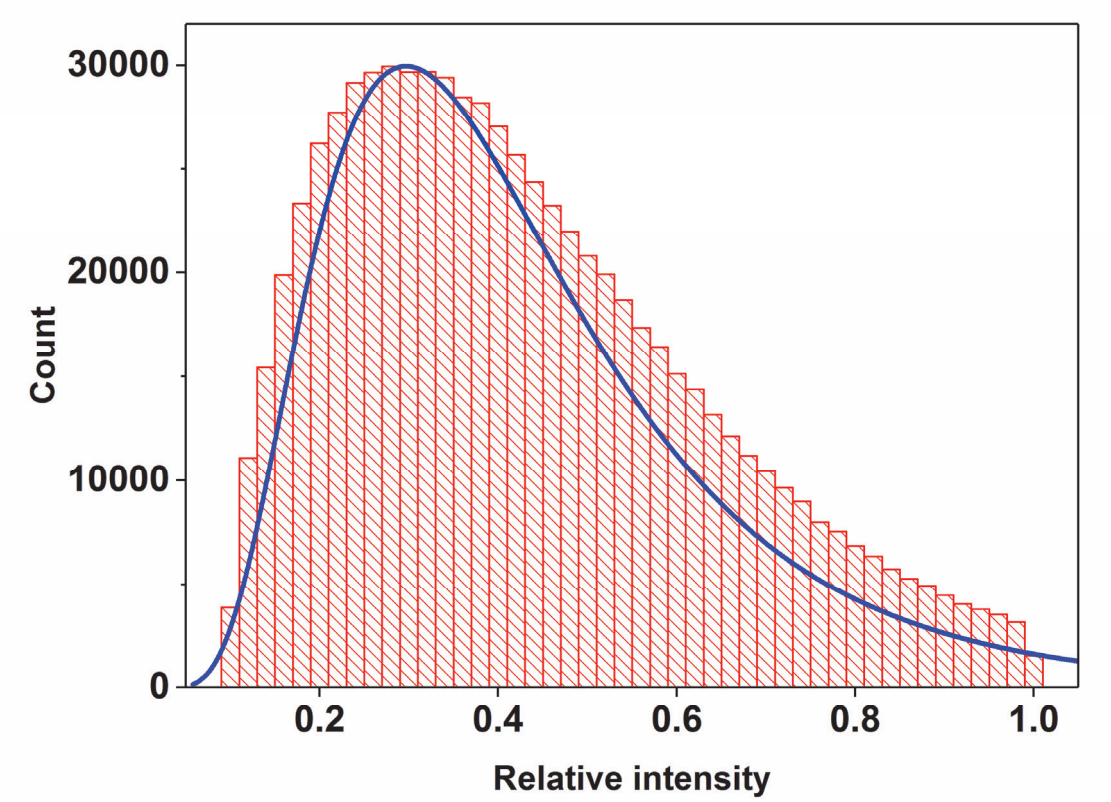
Supplementary Fig. 15. Twenty representative PSMs with mass shift of 34.006135 Da localized to His.



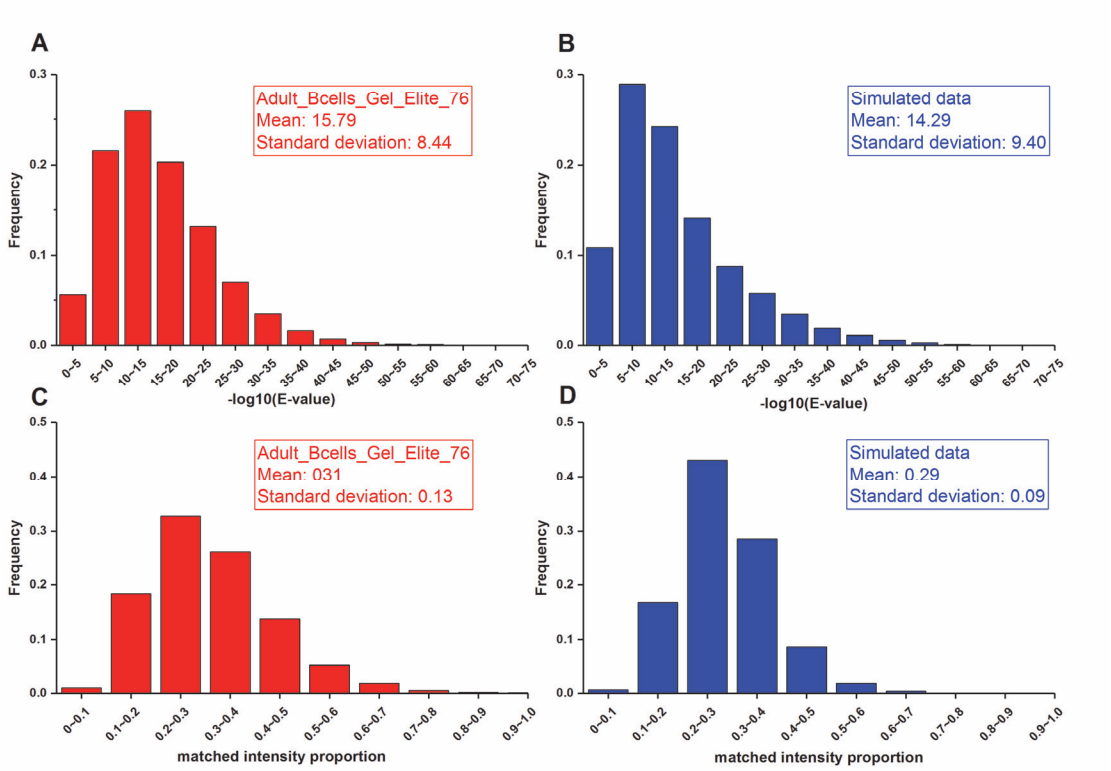
Supplementary Fig. 1. The FDRs of unmodified and differently modified peptides are likely to be different at the same score threshold. (A) Global FDR estimation is accurate on the whole set of PSMs of the simulated data set. The score threshold for 1% FDR is 20.27. (B-D) The group FDRs among differently modified and unmodified peptides are different at the same score thresholds. The group FDRs for individual modification groups are almost all equal to 0 at 1% global FDR score threshold of 20.27.



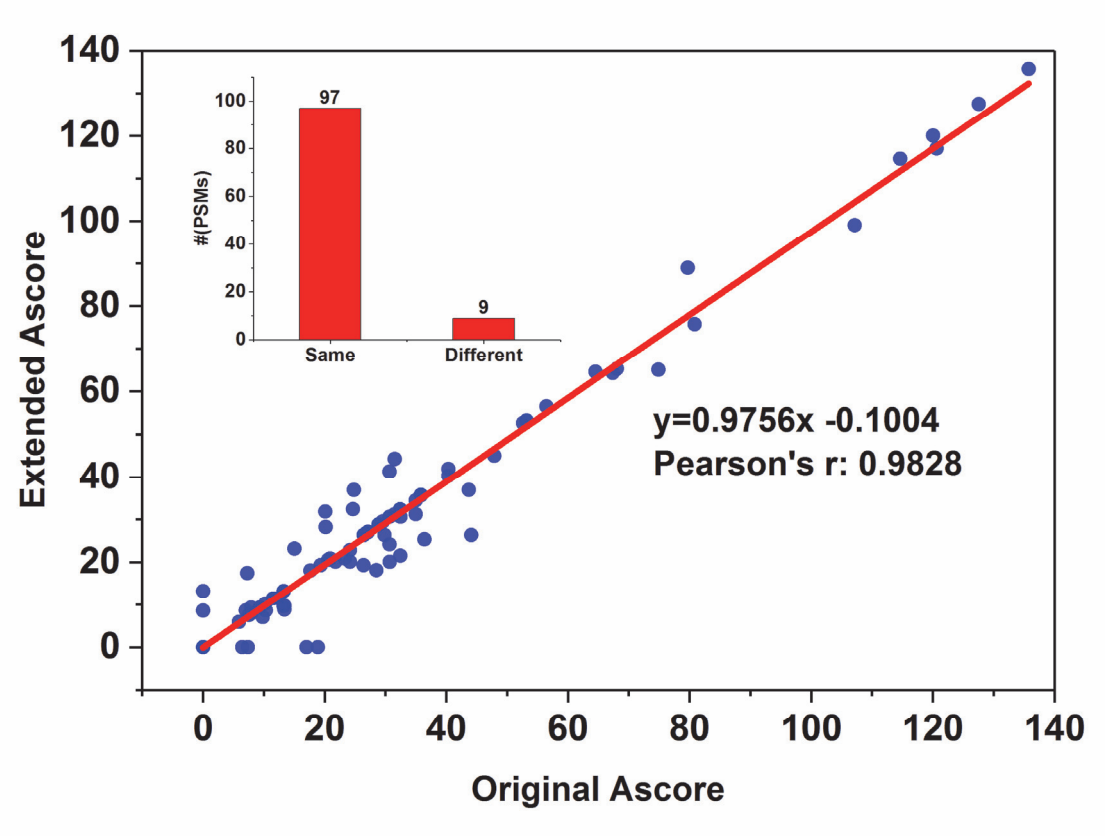
Supplementary Fig. 2. Example of $\gamma_k(x)$ in transfer FDR that was well fitted by linear regression. The data points come from the identifications with mass shifts in [15.5, 16.5] Da in one sample of adult adrenal gland of the draft map of human proteome (Adult_Adrenalgland_Gel_Elite_49).



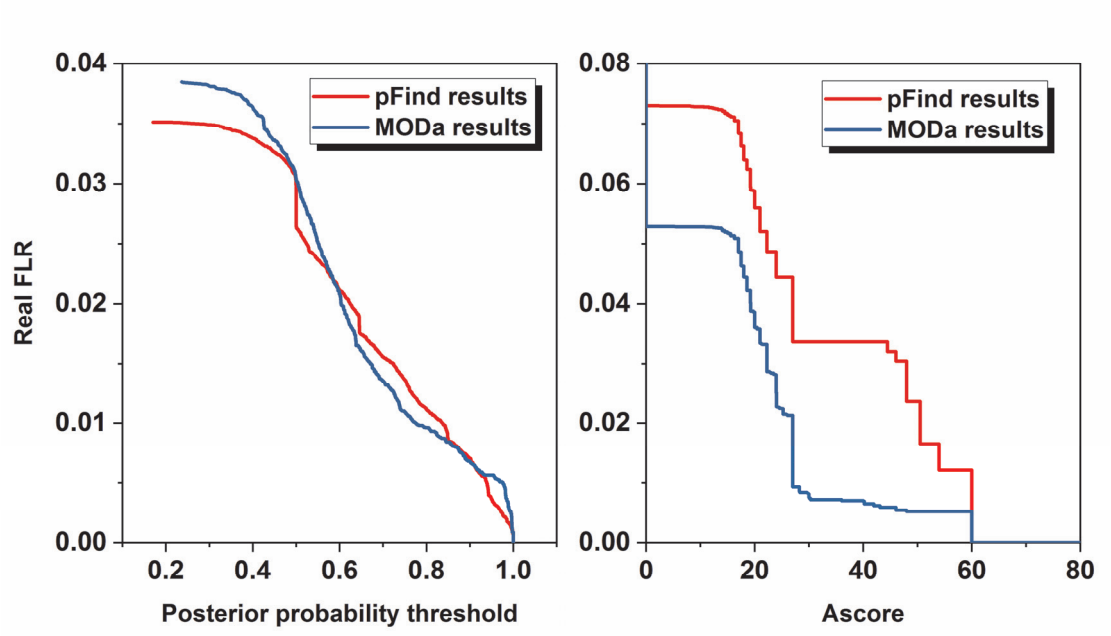
Supplementary Fig. 3. Example of relative intensities of matched peaks of unmodified PSMs, which were fitted using lognormal distribution. The data points come from the unmodified identifications from one sample of the adult adrenal gland of the draft map of human proteome (Adult_Adrenalgland_Gel_Elite_49).



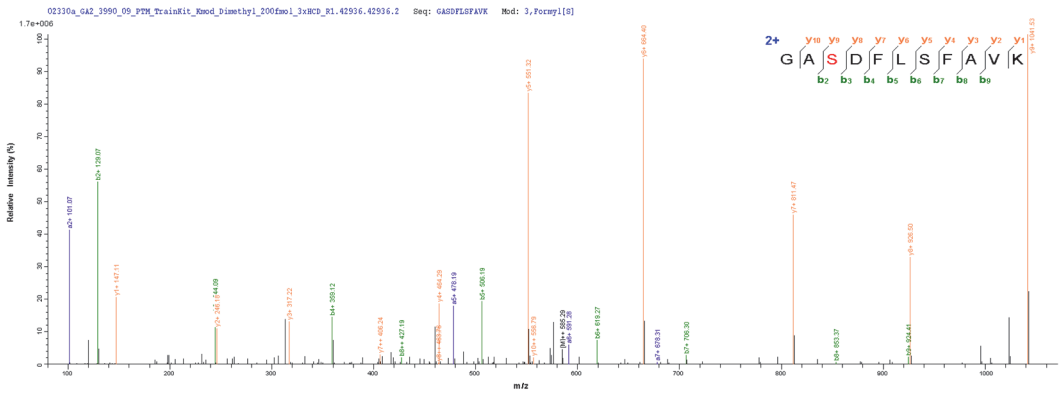
Supplementary Fig. 4. The evaluation of realness of the simulated data. Sub-figures A and B were the identification score distributions (*-log10(E-value)*) of the real data (one sample of the adult adrenal gland of the draft map of human proteome, Adult_Bcells_Gel_Elite_76) and simulated data, respectively. Sub-figures C and D were the matched intensity proportions (the sum of matched peaks' intensities over that of all peaks in one spectrum) of the real data and simulated data, respectively. As shown, these data between simulated and real spectra were very similar.



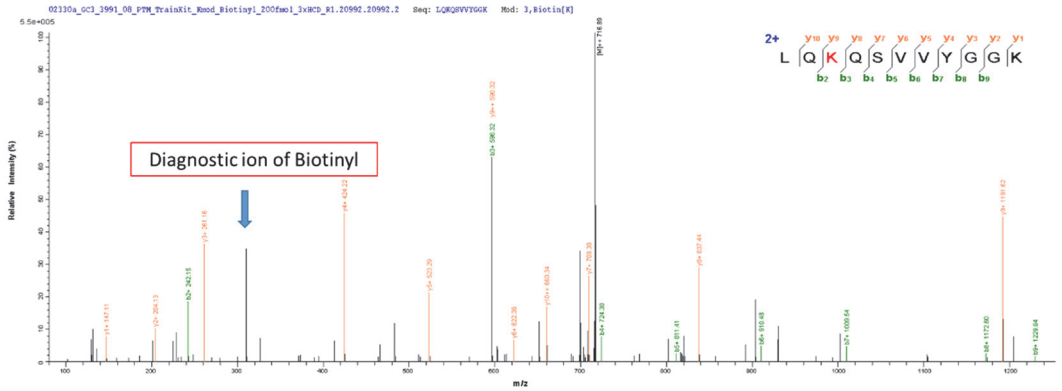
Supplementary Fig. 5. Comparison results of the original Ascore and our extended Ascore on phosphorylated PSMs. The scores of the two versions of Ascore software show strong positive correlation (Pearson's r is 0.9828) in the scatter plot. In fact, 55 (52%) scores were identical. As shown in the bar plot, 97 (92%) phosphorylations were localized to the same sites on peptides. The small difference between the two versions of Ascore was probably due to the preprocessing steps before localization.



Supplementary Fig. 6. The real FLRs are plotted against the (A) PTMiner posterior probability and (B) Ascore for the open search results of the simulated data set. The posterior probability thresholds of 1% FLR for pFind and MODa results were similar (0.837 for pFind and 0.775 for MODa), but the Ascore thresholds of 1% FLR for pFind and MODa results were quite different (60 for pFind and 27 for MODa).

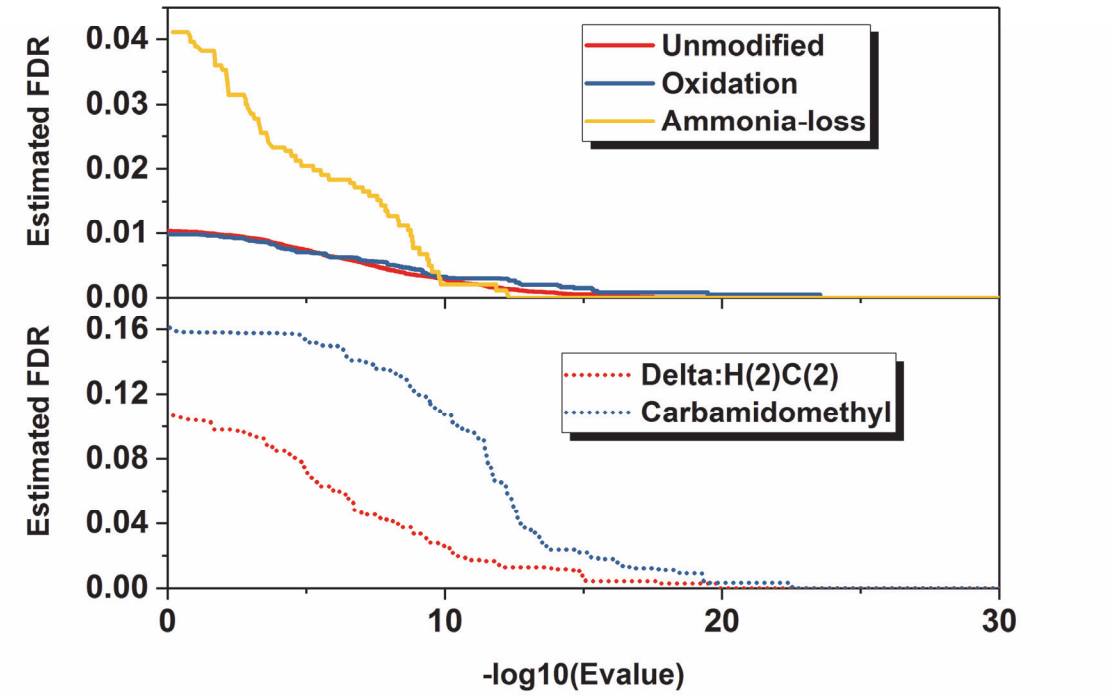


(A)

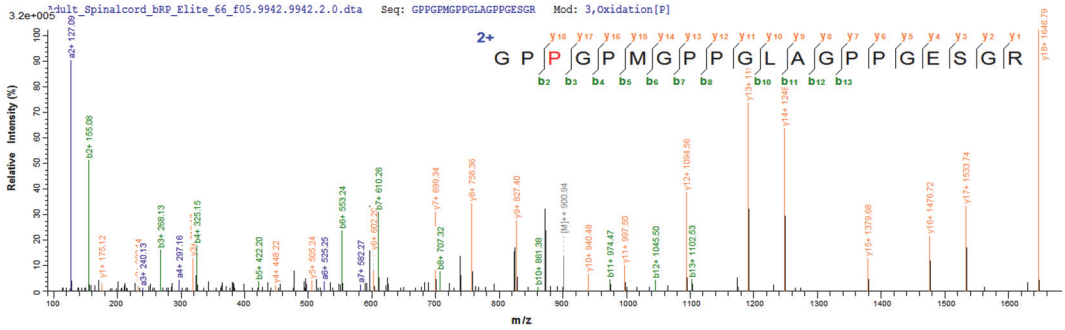


(B)

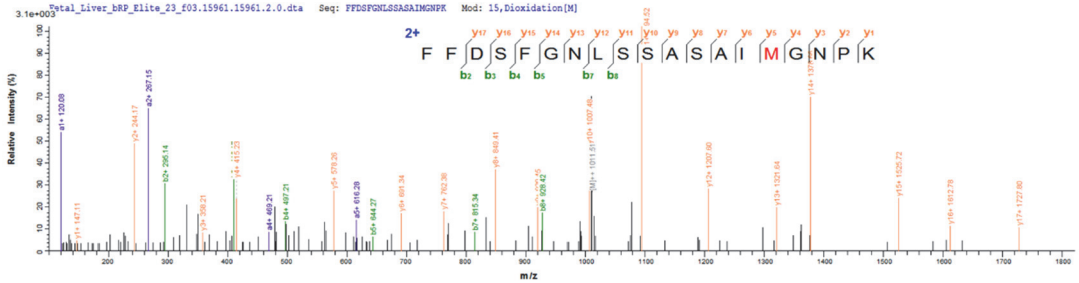
Supplementary Fig. 7. Examples of identified peptides that do not belong to the chemically synthesized modified peptides in the ProteomeTools project. (A) A spiked-in quality control peptide (B) A byproduct $LQK_{biotin}QSVVYGGK$ from the expected synthesized product $TLQK_{biotin}QSVVYGGK$. The identified peptide by pFind was TLQKQSVVYGGK, and the mass shift was 125.03343 Da, unequal to the mass of the expected biotinylation modification (226.077598 Da). However, if we removed the first T (101.04768 Da) from the identified peptide, then the mass shift would be 226.08111 Da, matching to the mass of a biotinyl group (the mass error is 0.003512 Da). We also found the diagnostic ion of the modification biotinyl, i.e. the ion of 310.1584 Da.



Supplementary Fig. 8. The FDRs of different modifications are different at the same score threshold for the data set of draft map of human proteome. Here, the FDRs were estimated separately for different modification groups using the conventional target-decoy approach. The spectra in this example came from one sample of adult liver (bRP_Elite_82) from the draft map of human proteome.

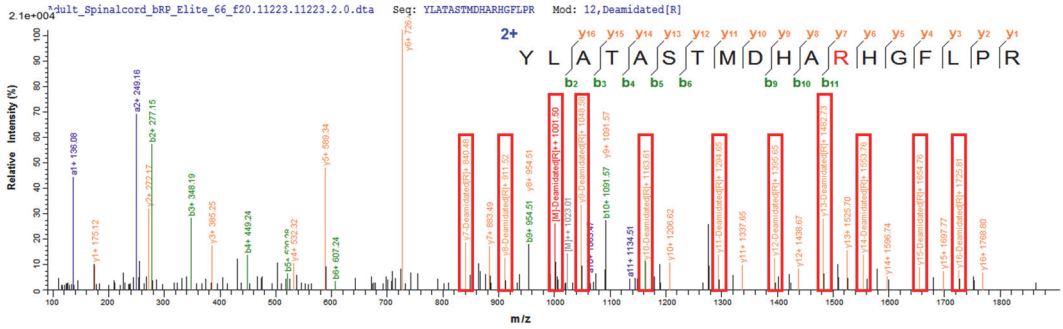


A. Oxidation [P]

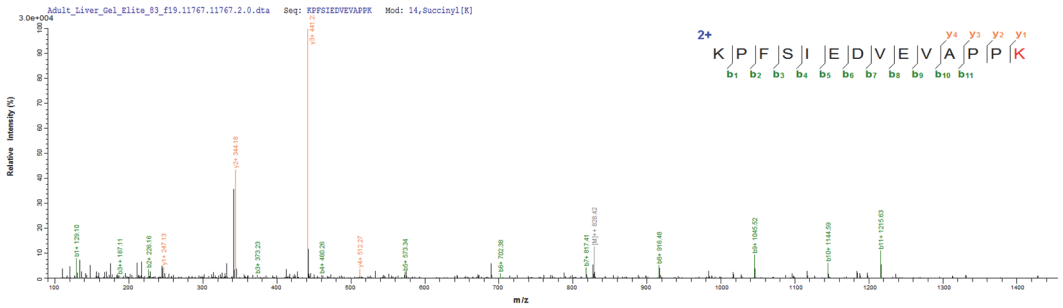


B. Di-oxidation [M]

Supplementary Fig. 9. The two examples of post-translational modifications (A is for Oxidation [P] and B is for Di-oxidation [M]) identified and localized with high confidence from the draft map of human proteome.

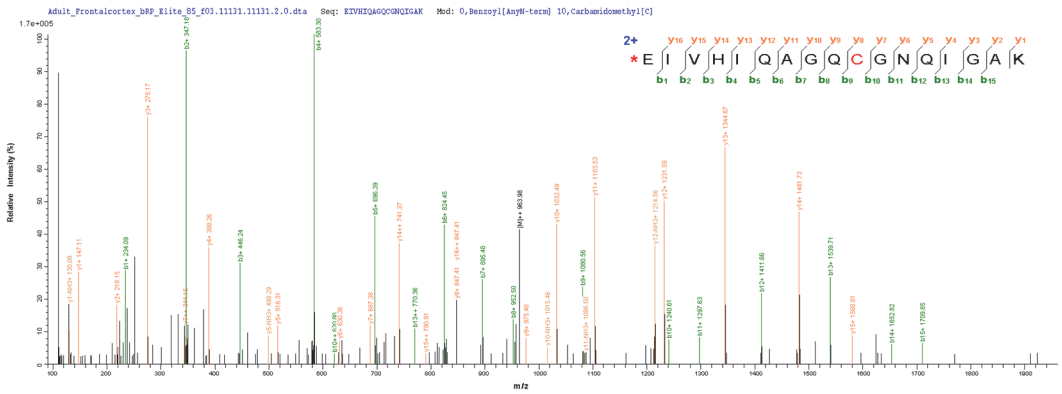


A. Deamidated [R]

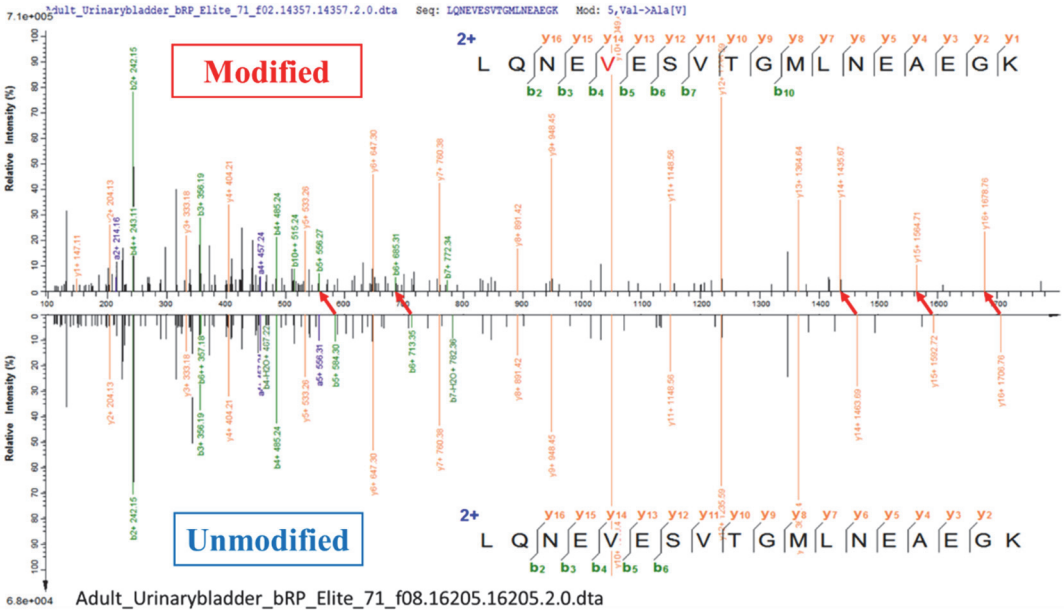


B. Succinyl [K]

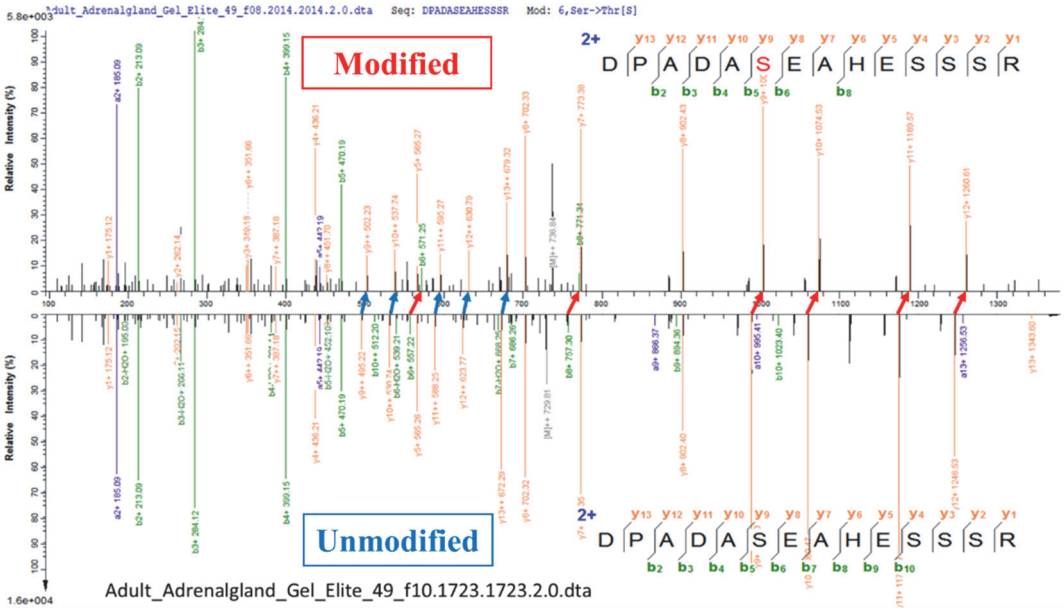
Supplementary Fig. 10. Two examples of MS/MS spectra of peptides with deamidated at Arg and succinlation at Lys. (A) MS/MS spectrum of a peptide with a deamidation at arginine residue is shown. The peaks in the red block are the peaks of neutral loss of isocyanic acid (HNCO, accurate mono mass is 43.005814 Da). This modified arginine residue comes from the 159-th amino acid of myelin basic protein (MBP), which has 14 *in vivo* arginine residues likely modified by deamidation. (B) MS/MS spectrum of a peptide with a modification of succinyl at lysine.



Supplementary Fig. 11. One PSM example for benzoylation modification.

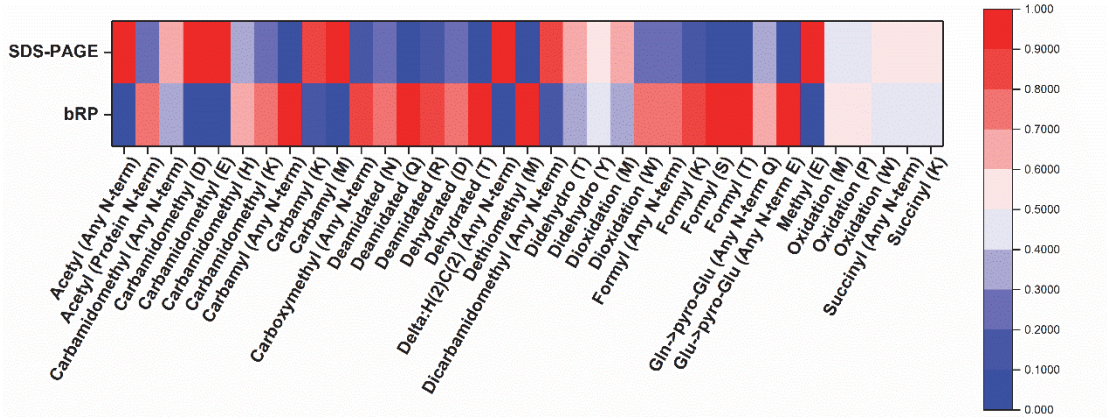


A. rs16967510

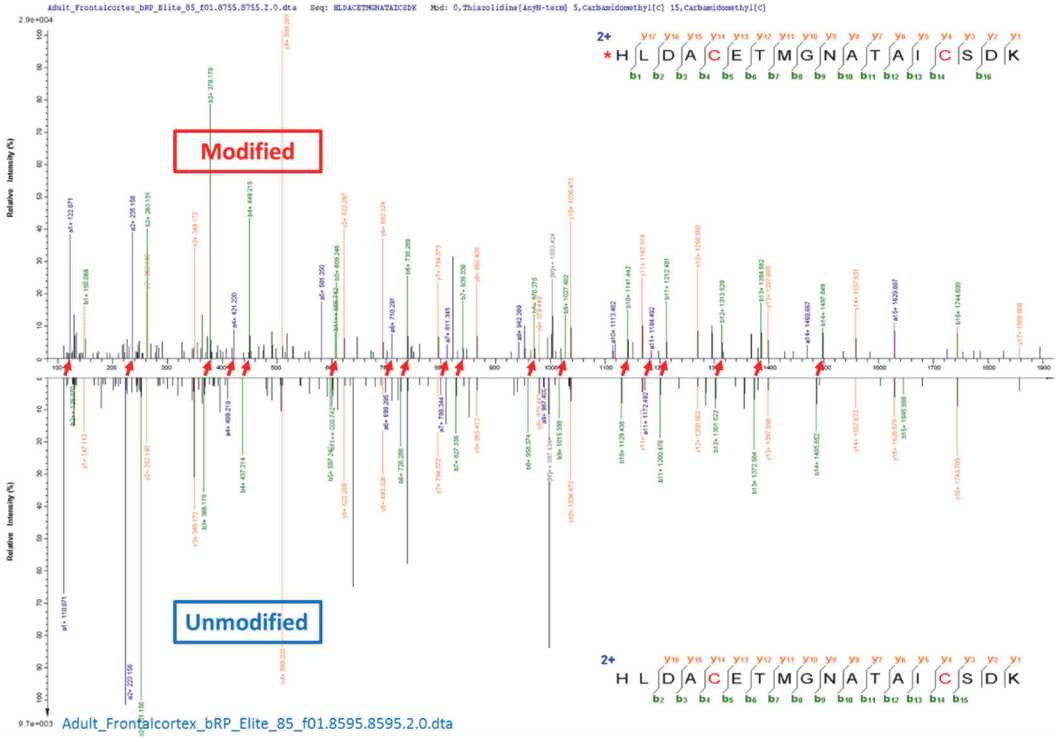


B. rs6085324

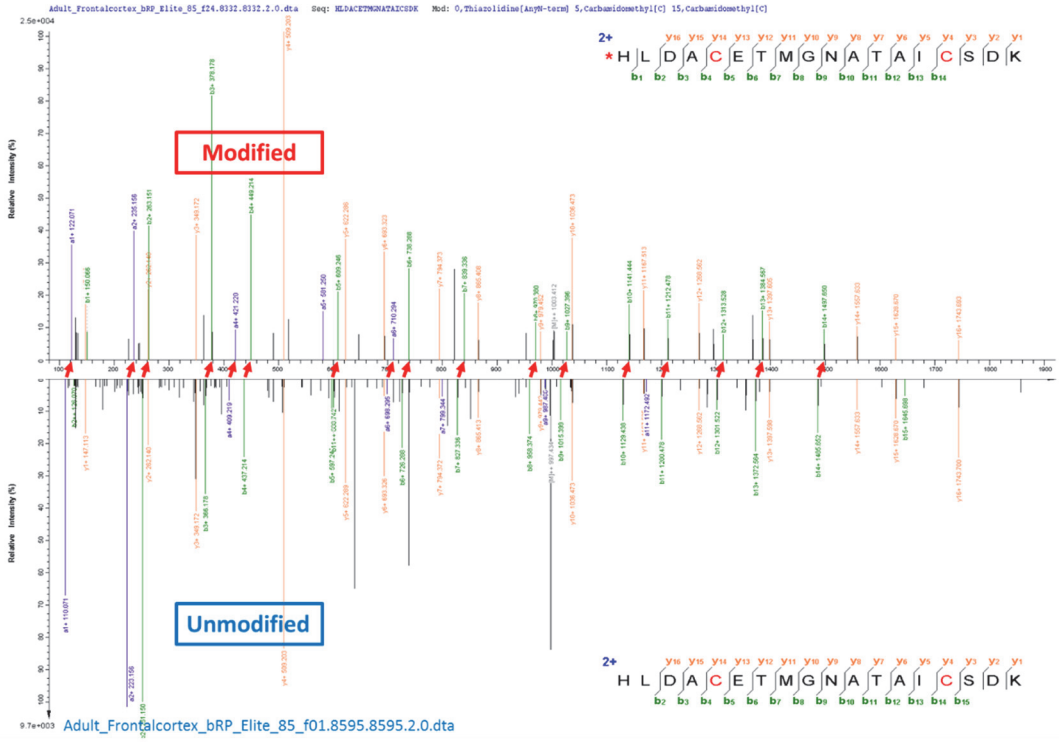
Supplementary Fig. 12. Two examples of MS/MS spectra of peptides with single amino acid variations (SAVs) annotated in the UniProt database. (A) and (B) correspond to SAVs of rs16967510 (V1289A of Myosin-11 protein) and rs6085324 (S93T of Secretogranin-1 protein), respectively. MS/MS spectra of variant (top) and normal (bottom) forms of two peptides are shown and compared. Arrows between spectra (red and blue for singly and doubly charged ions, respectively) represent mass shift. The spectra of the variant peptides are very similar to those of the normal peptides in terms of their fragmentation pattern even for these ions undergoing mass migration.



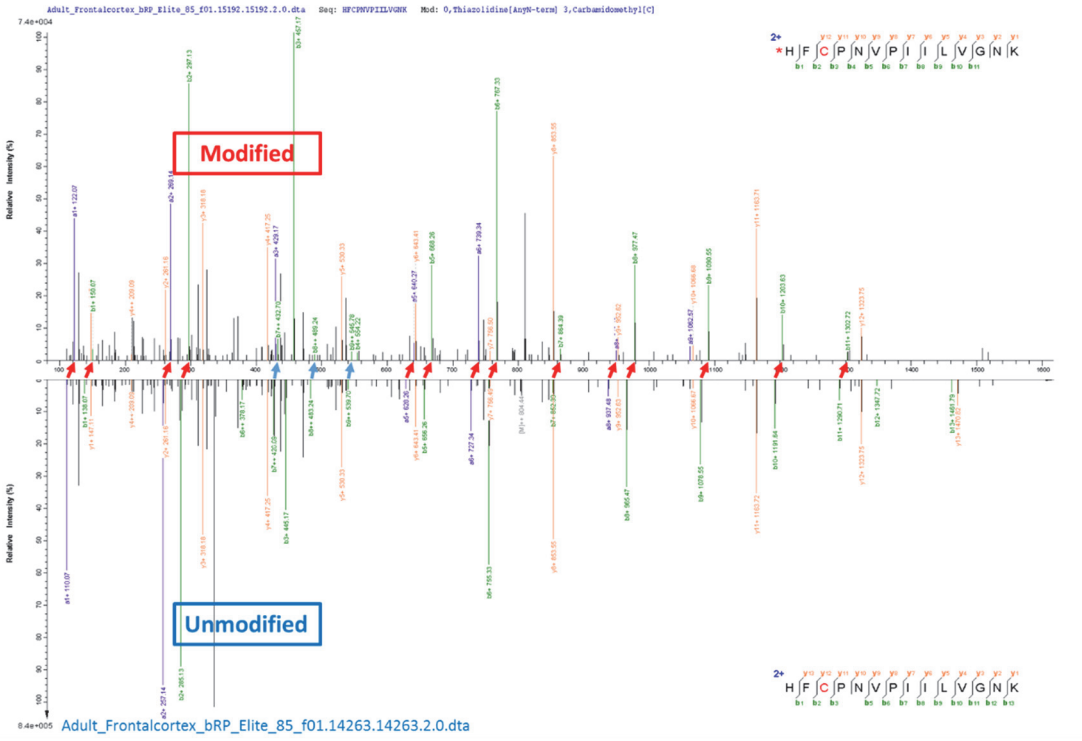
Supplementary Fig. 13. Some fully annotated modifications show strong differences between sample preparation methods (SDS-PAGE and bRP). The heat map was plotted as follows. First, we counted the numbers of PSMs for every modification type in each of the two sample preparations, then these numbers were divided by the total spectral numbers of the corresponding situations, and finally normalization was performed such that the sum of the two situations for each modification was 1.


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f01.8595.8595.2.0.dta

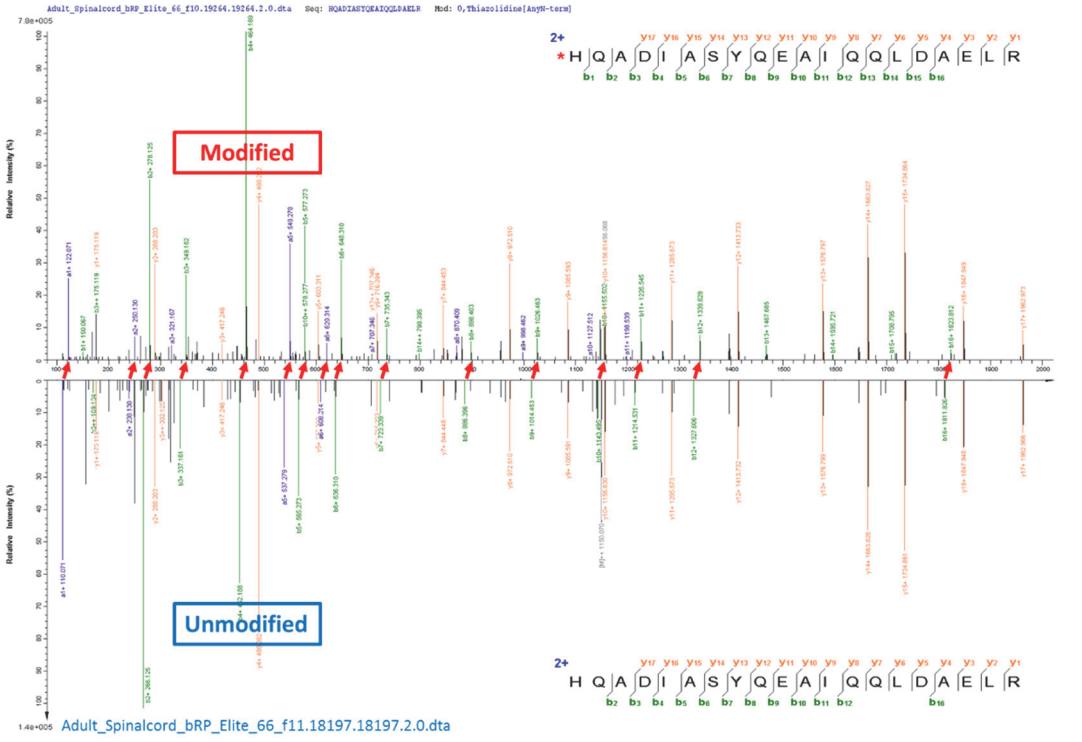
A-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f01.8595.8595.2.0.dta

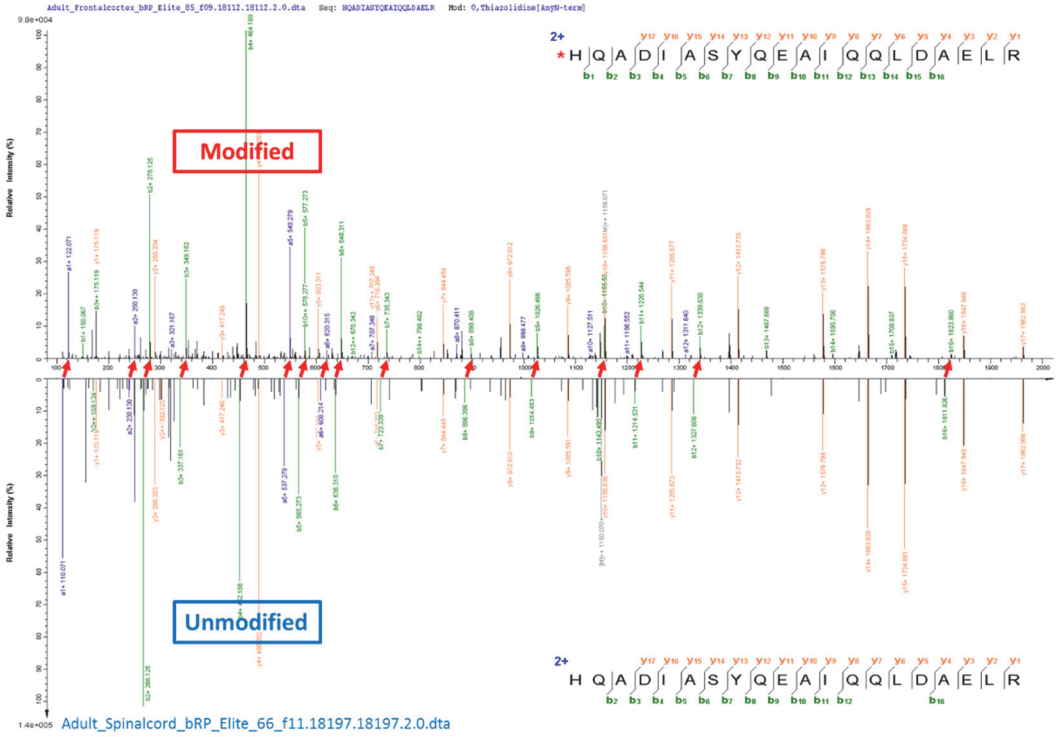
A-2


Modified
Unmodified
Relative Intensity (%)
Adult_Frontalcortex_bRP_Elite_85_f01.14263.14263.2.0.dta

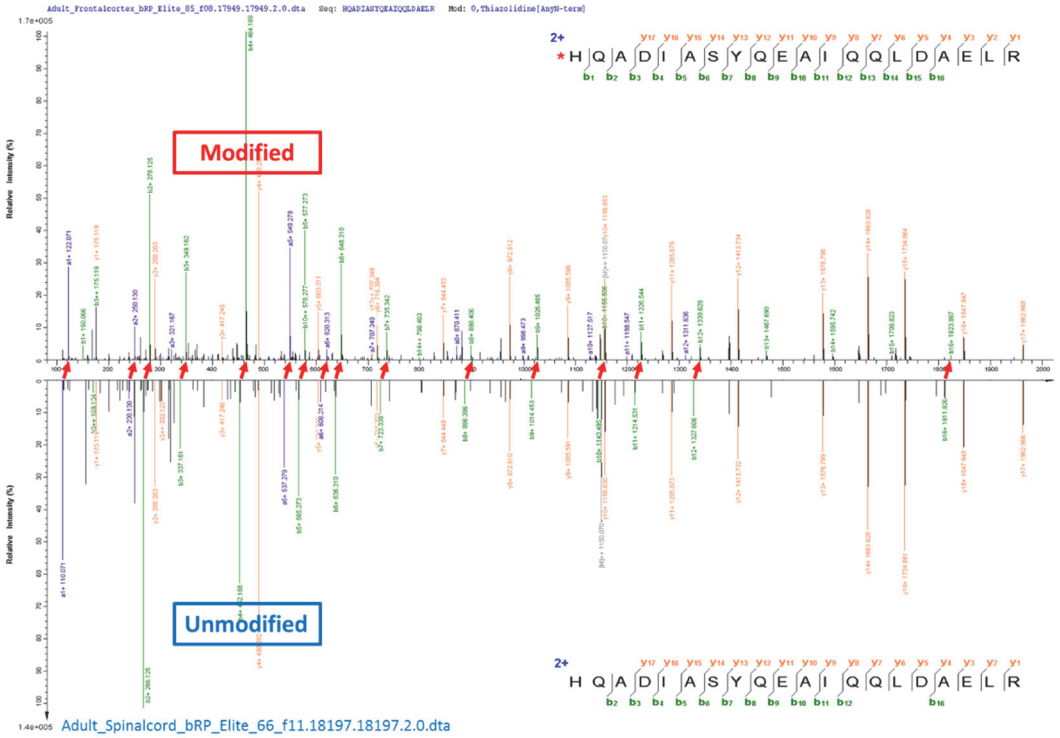
B-1


Modified
Unmodified
Relative Intensity (%)
Adult_Spinalcord_bRP_Elite_66_f11.18197.18197.2.0.dta

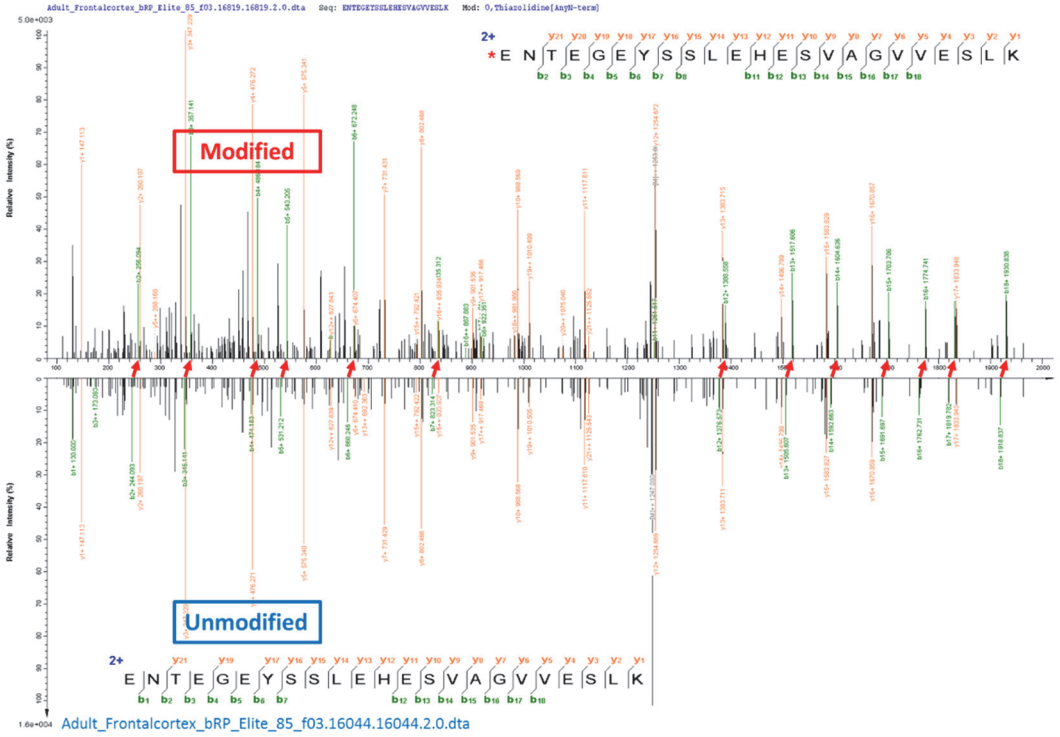
C-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f11.18197.18197.2.0.dta

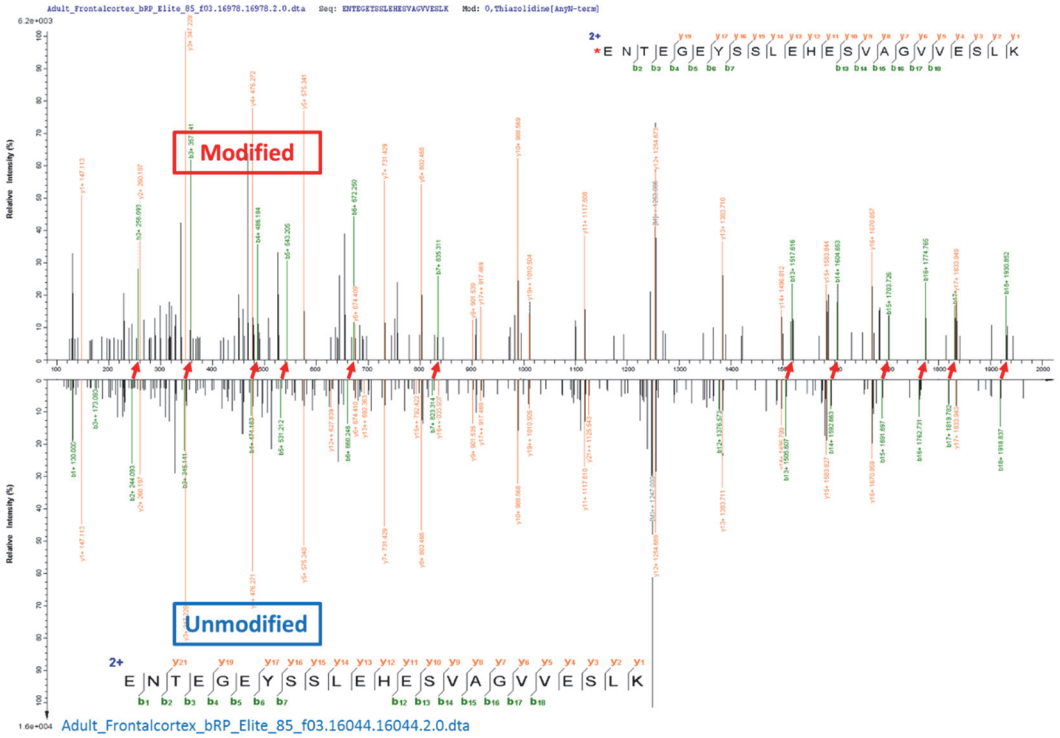
C-2


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f11.18197.18197.2.0.dta

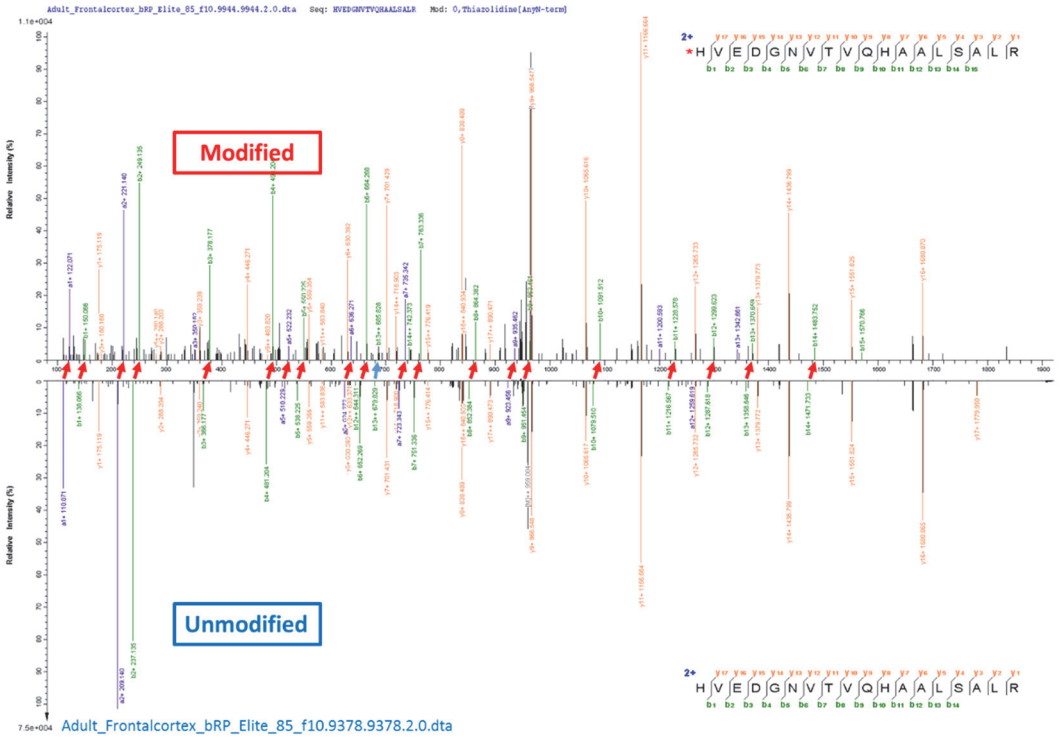
C-3


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f03.16044.16044.2.0.dta

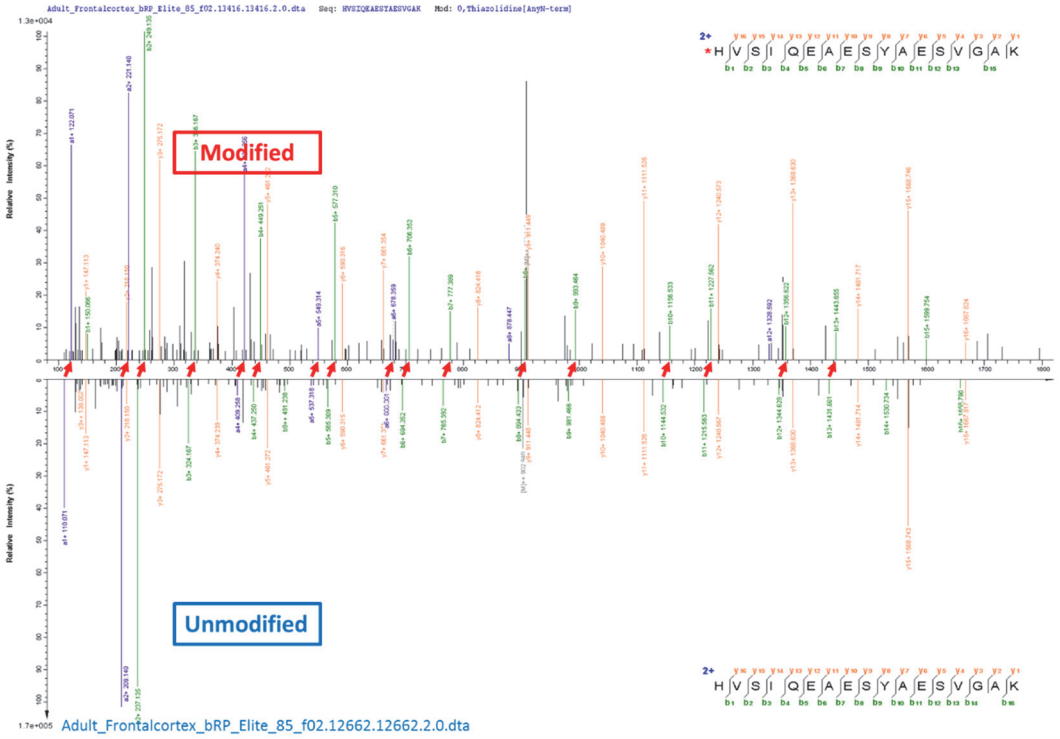
D-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f03.16044.16044.2.0.dta

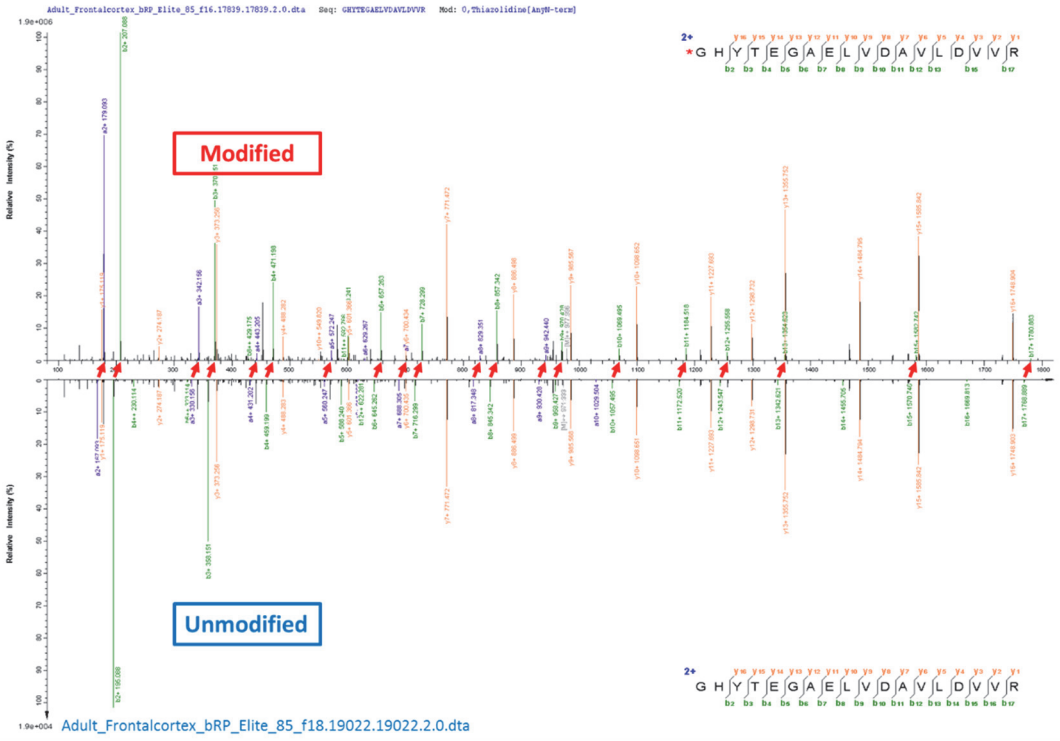
D-2


Modified
Unmodified
Relative Intensity (%)
Adult_Frontalcortex_bRP_Elite_85_f10.9378.9378.2.0.dta

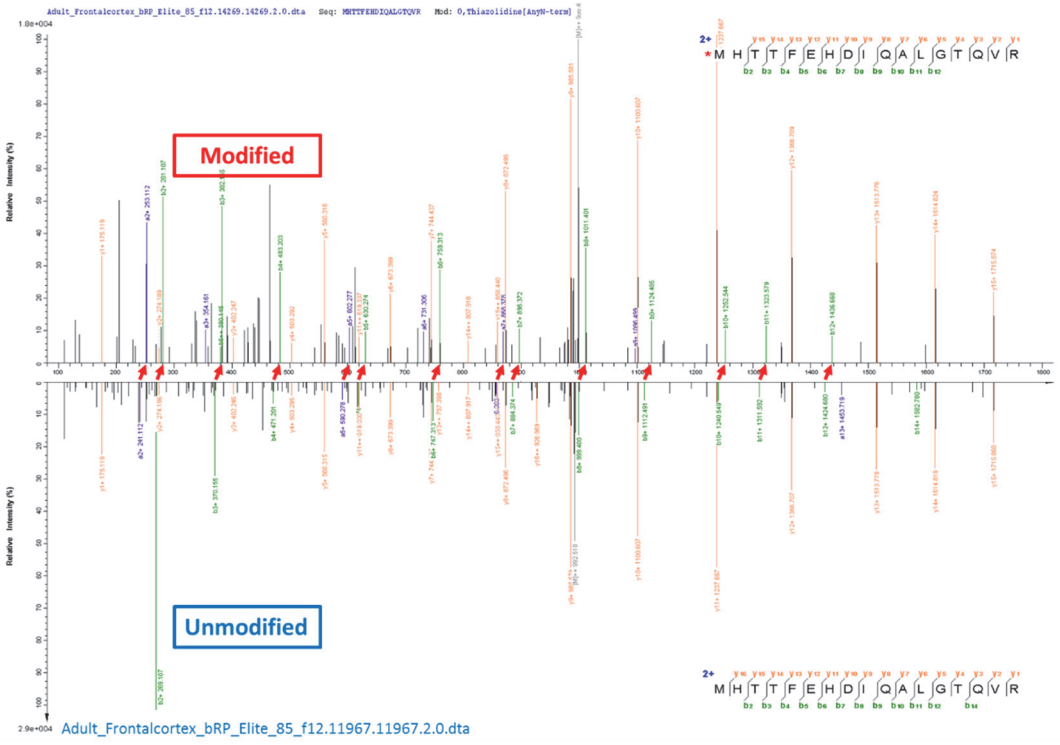
E-1


Modified
Unmodified
Relative Intensity (%)
Adult_Frontalcortex_bRP_Elite_85_f02.12662.12662.2.0.dta

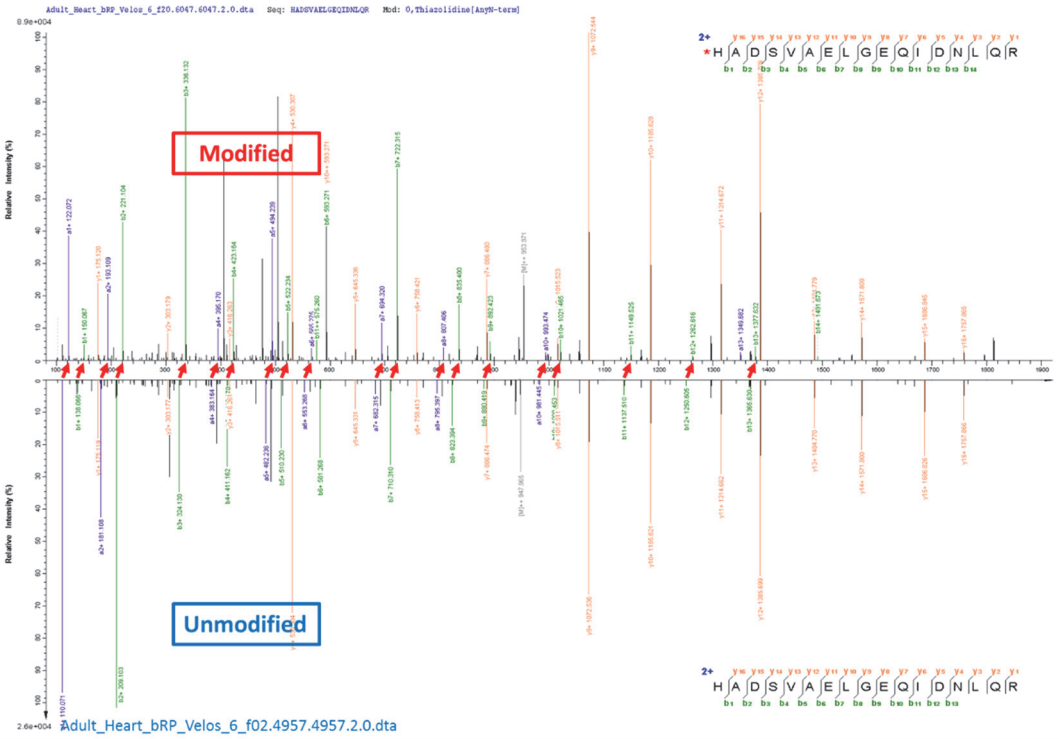
F-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f18.19022.19022.2.0.dta

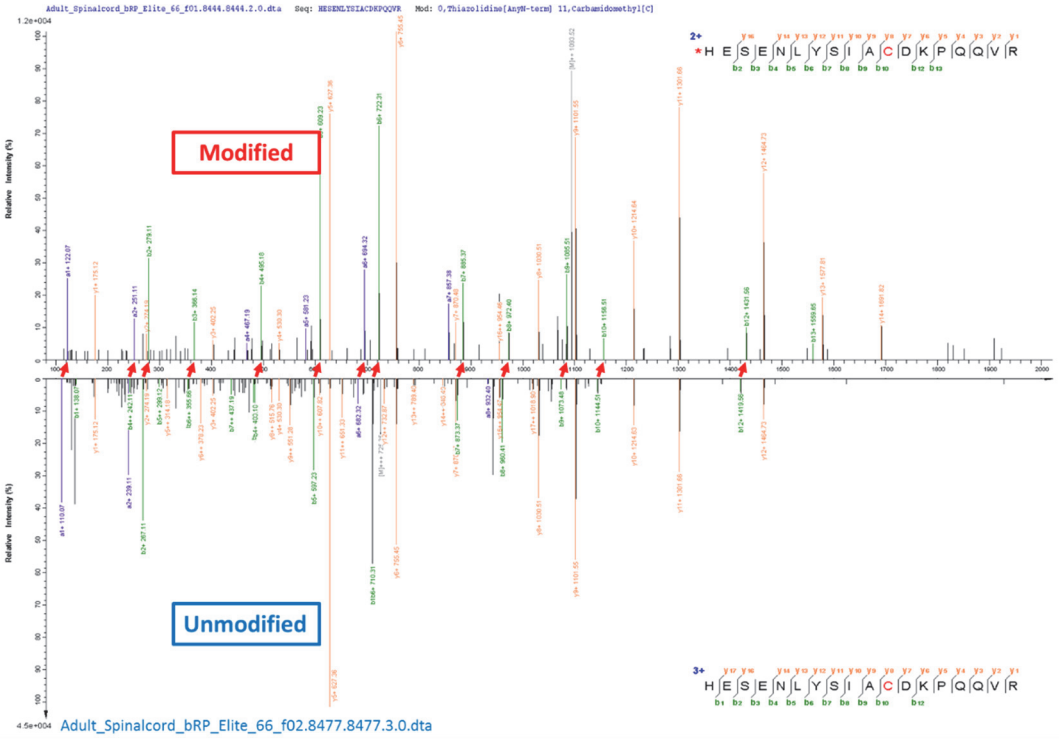
G-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f12.11967.11967.2.0.dta

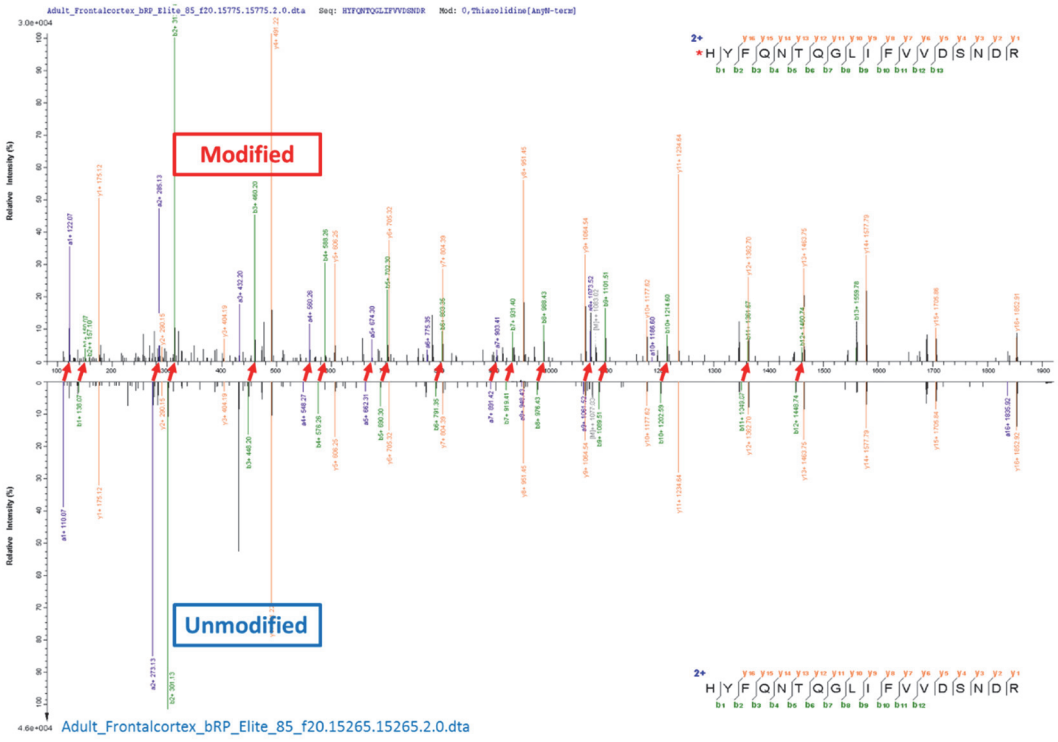
H-1


Modified
Unmodified
Adult_Heart_bRP_Velos_6_f02.4957.4957.2.0.dta

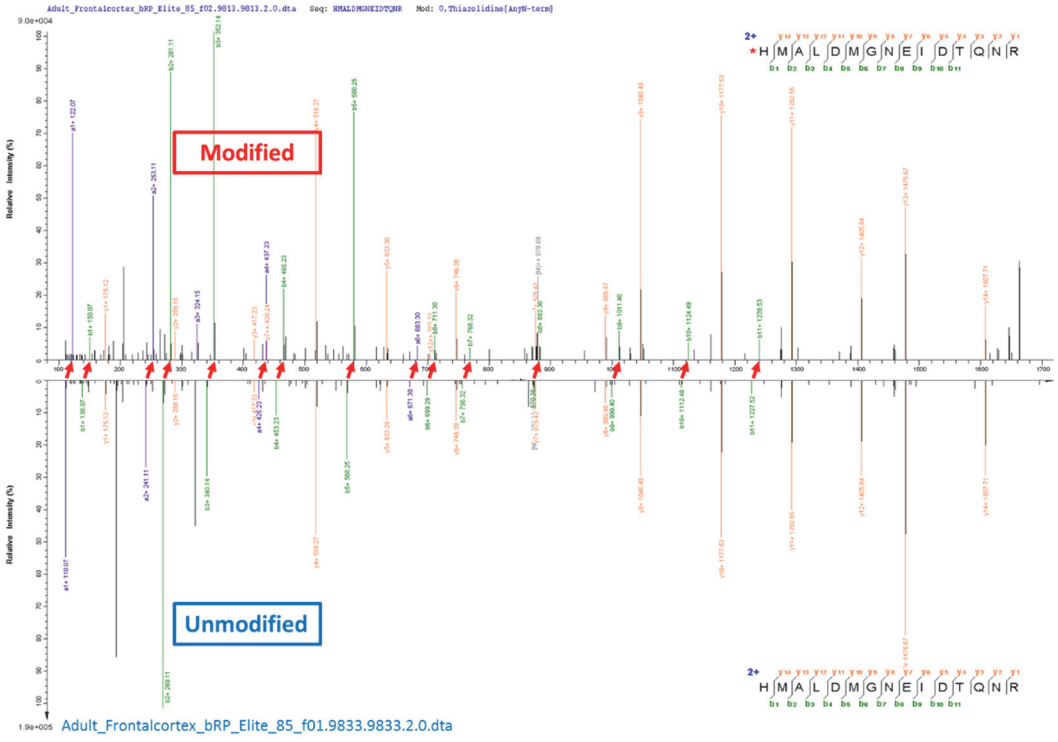
I-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f02.8477.8477.3.0.dta

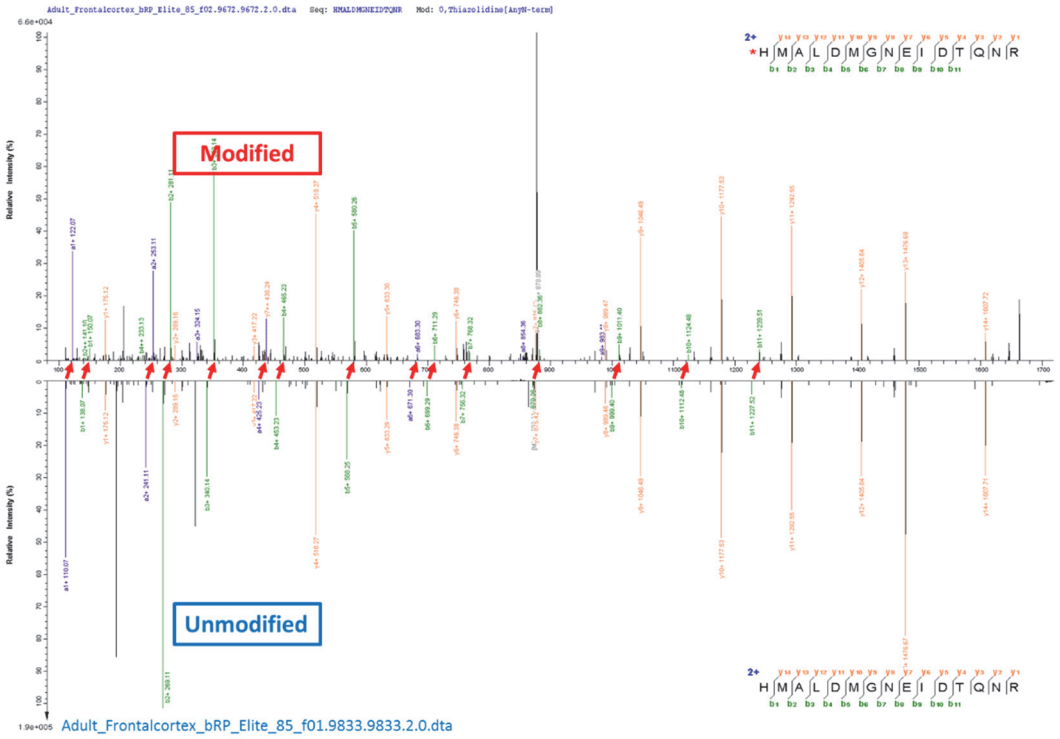
J-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f20.15265.15265.2.0.dta

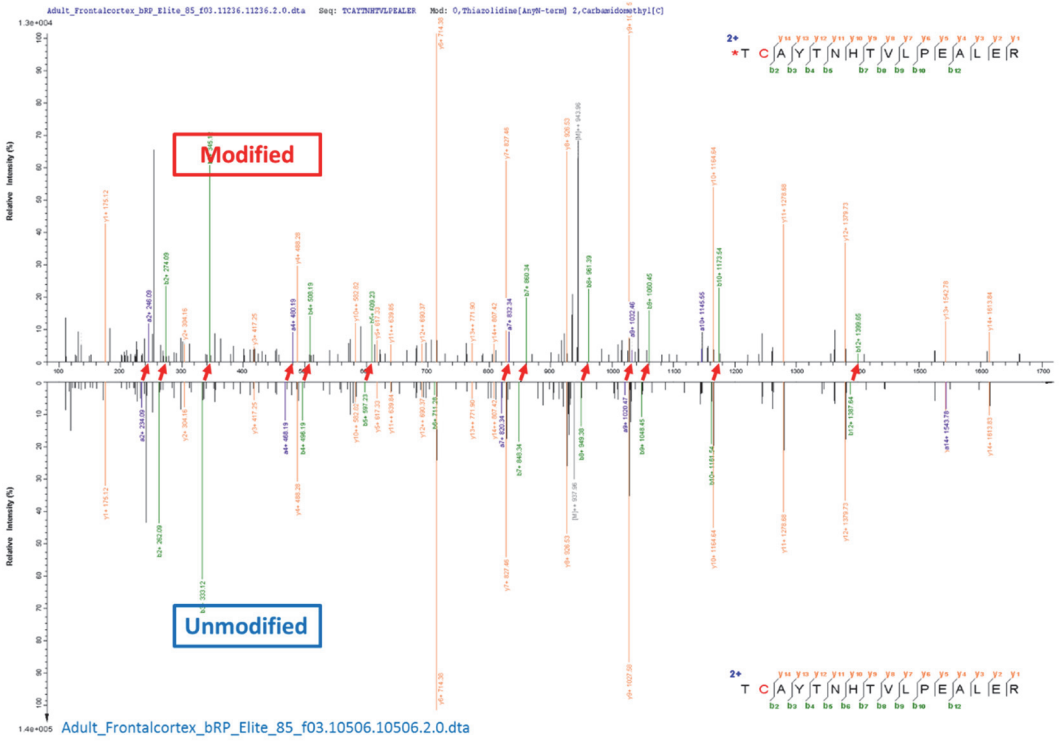
K-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f01.9833.9833.2.0.dta

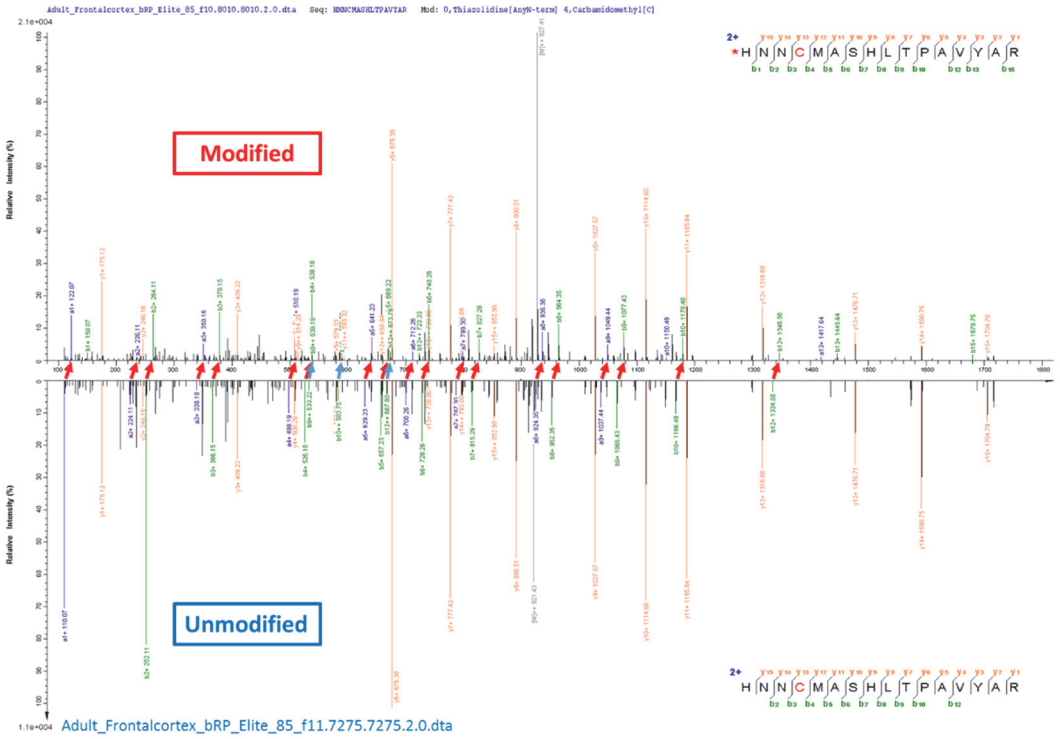
L-1


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f01.9833.9833.2.0.dta

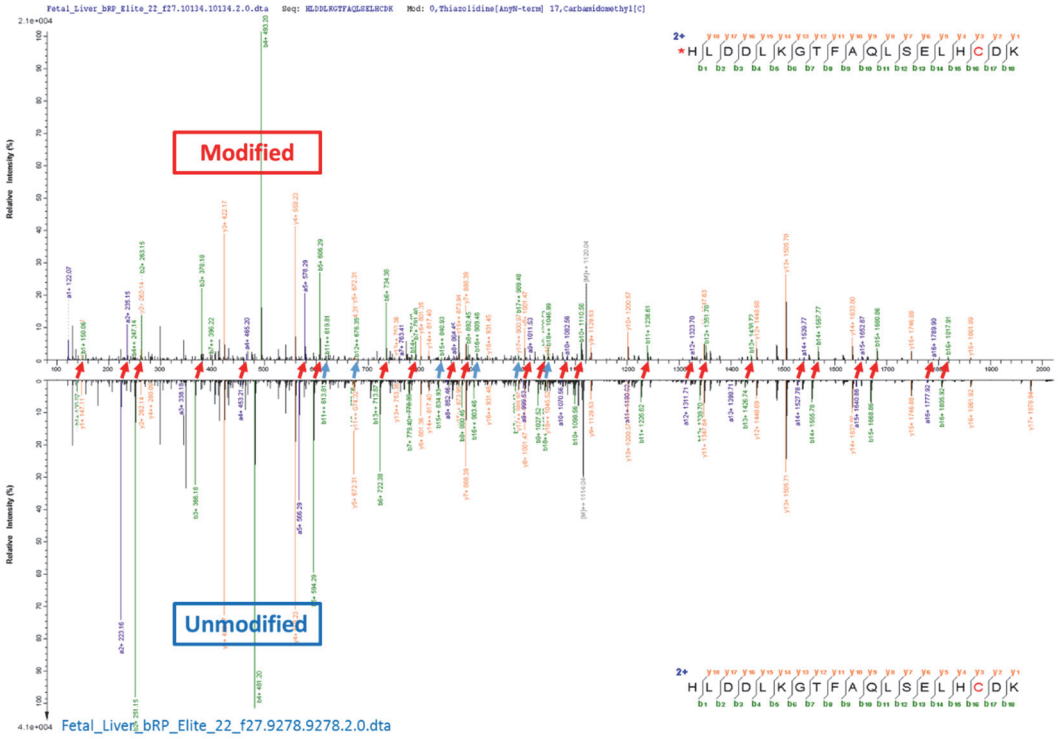
L-2


Modified
Unmodified
Adult_Frontalcortex_bRP_Elite_85_f03.10506.10506.2.0.dta

M-1



N-1



O-1

Supplementary Fig. 14. Twenty representative PSMs with mass shift of 12.000000 Da localized to peptide N-terminal. MS/MS spectra of modified (top) and unmodified (bottom) forms of the peptide

are shown and compared. Red and blue arrows between two spectra indicate the shift of fragment ion peaks with one and two, respectively. The indexes of A-O indicate different peptide sequences.

A (2): HLDACETMGNATAICSDK

B (1): HFCPNVPIILVGNK

C (3): HQADIASYQEAIQQLDAELR

D (2): ENTEGEYSSLEHESVAGVVESLK

E (1): HVEDGNVTVQHAALSALR

F (1): HVSIQEAESYAESVGAK

G (1): GHYTEGAELVDAVLDVVR

H (1): MHTTFEHDIQALGTQVR

I (1): HADSVAELGEQIDNLQR

J (1): HESENLYSIACDKPQQVR

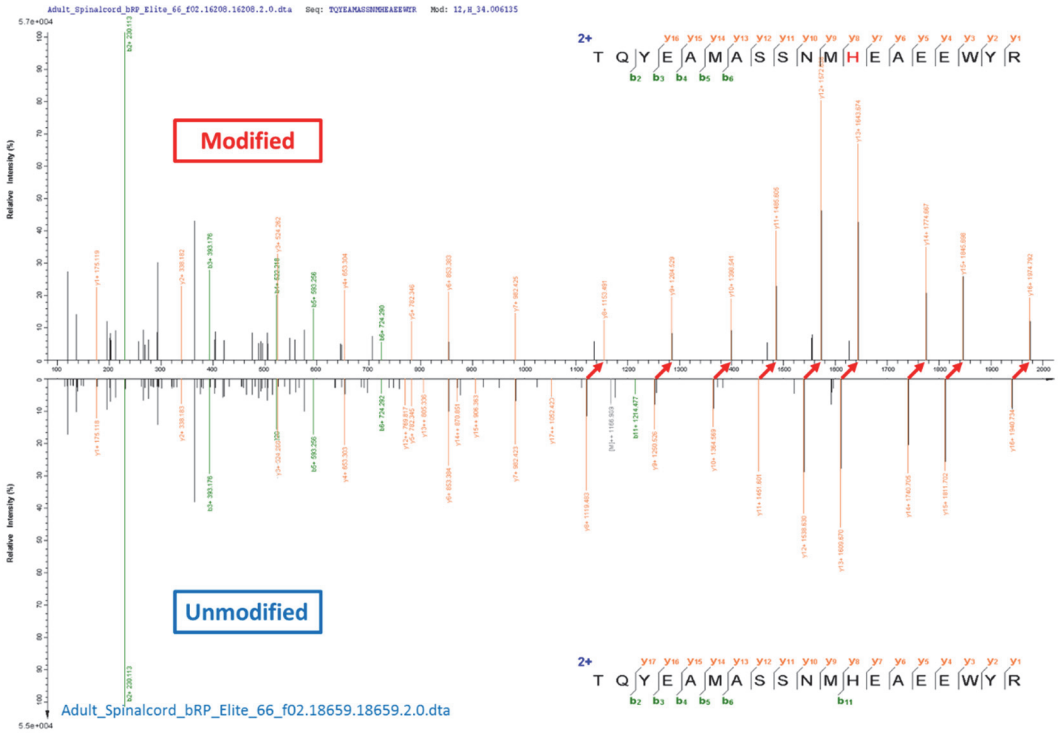
K (1): HYFQNTQGLIFVVDSNDR

L (2): HMALDMGNEIDTQNR

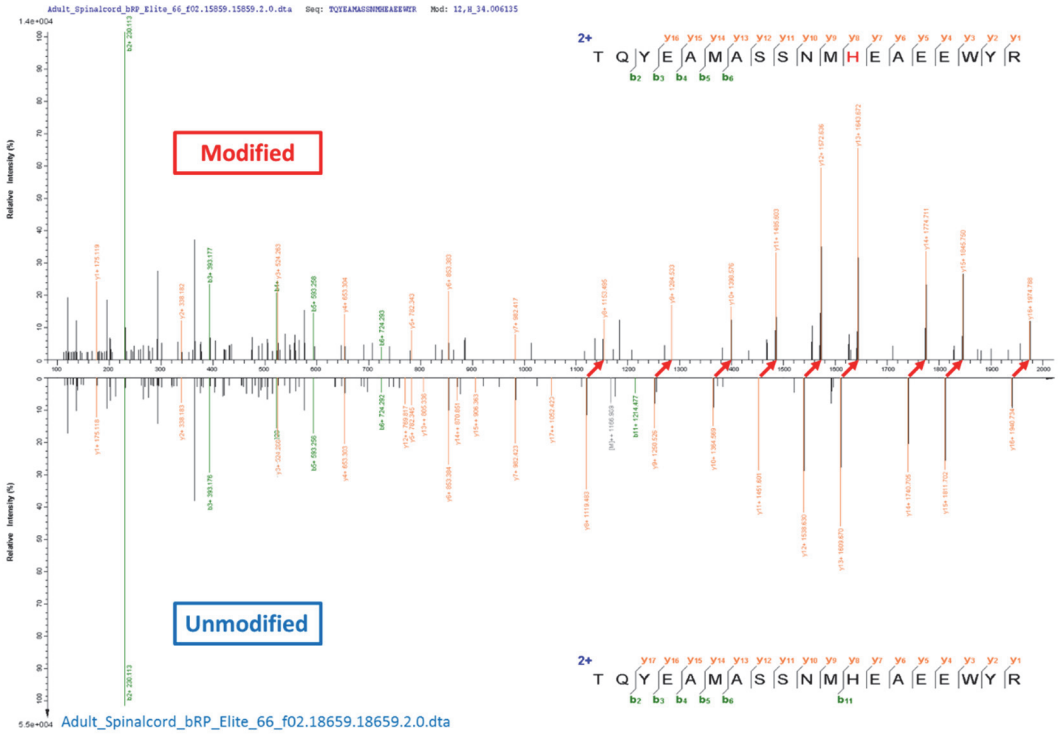
M (1): TCAYTNHTVLPEALER

N (1): HNNCMASHLTPAVYAR

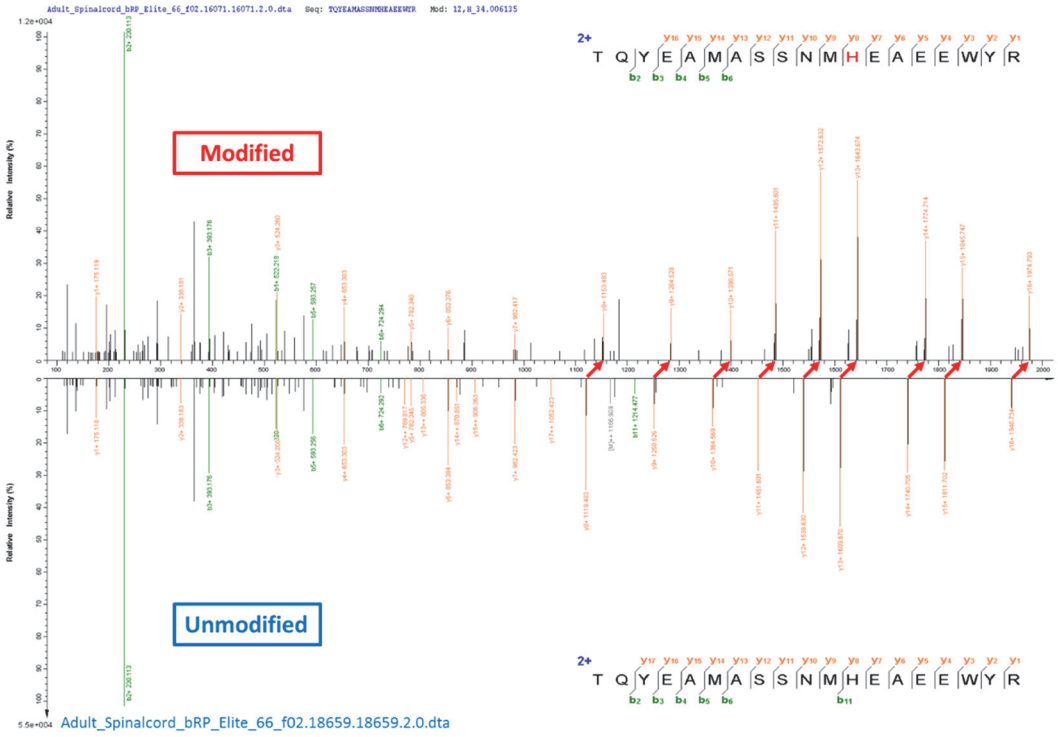
O (1): HLDDLKGTFAQLSELHCDK


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f02.18659.18659.2.0.dta
T Q Y E A M A S S N M H E A E E W Y R

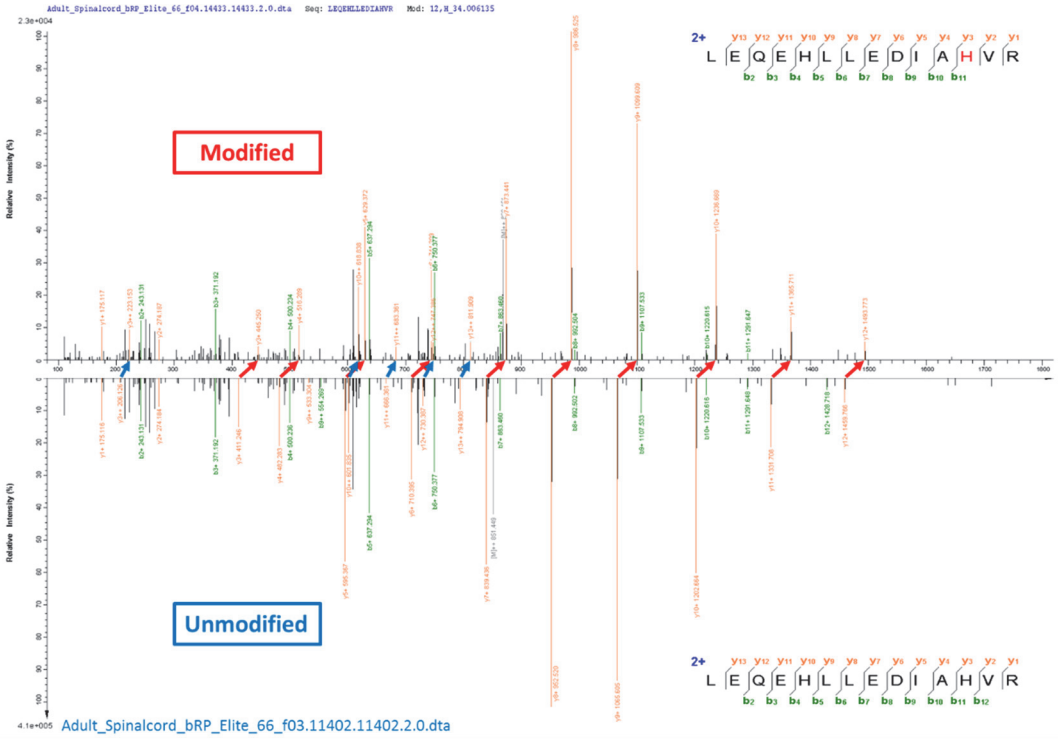
A-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f02.18659.18659.2.0.dta
T Q Y E A M A S S N M H E A E E W Y R

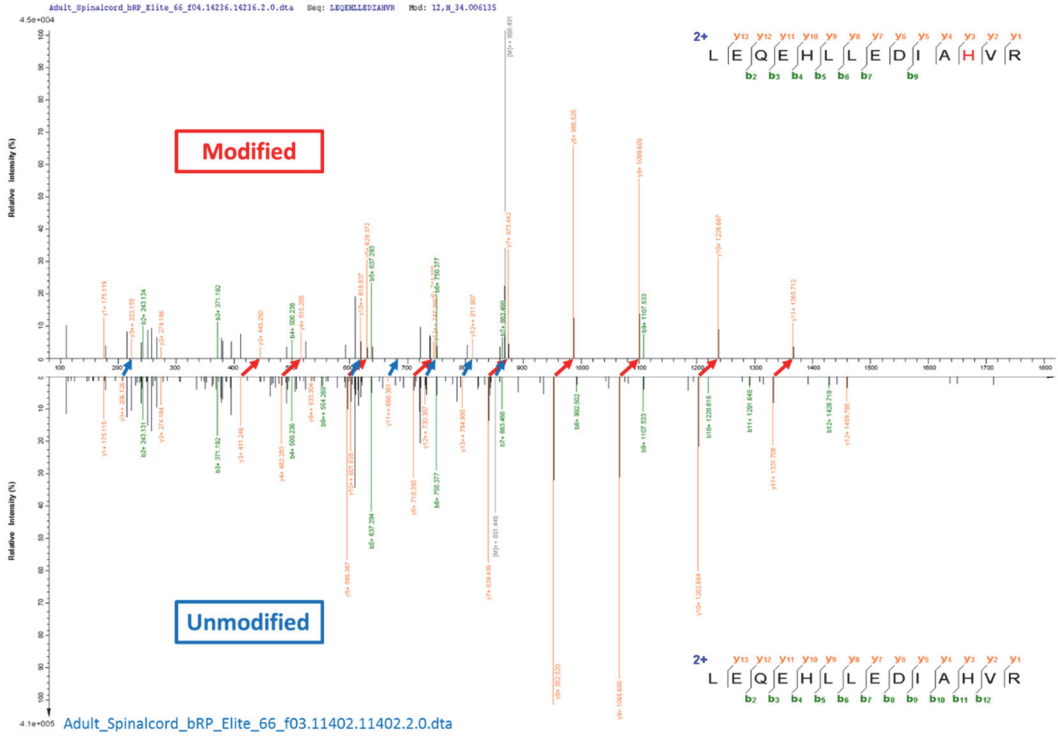
A-2


Modified
Unmodified
Relative Intensity (%)
2+
T Q Y E A M A S S N M H E A E E W Y R
b2 b3 b4 b5 b6
b11
Adult_Spinalcord_bRP_Elite_66_f02.18659.18659.2.0.dta

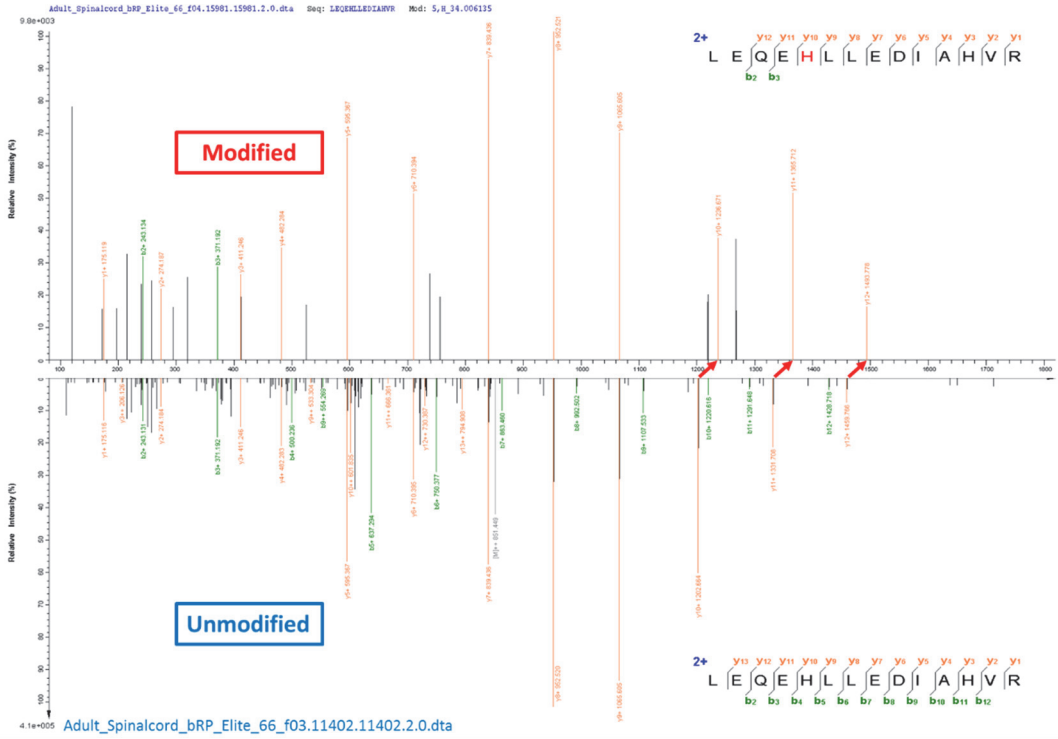
A-3


Modified
Unmodified
Relative Intensity (%)
2+
L E Q E H L L E D I A H V R
b2 b3 b4 b5 b6 b7 b8 b9 b10 b11
b12
Adult_Spinalcord_bRP_Elite_66_f03.11402.11402.2.0.dta

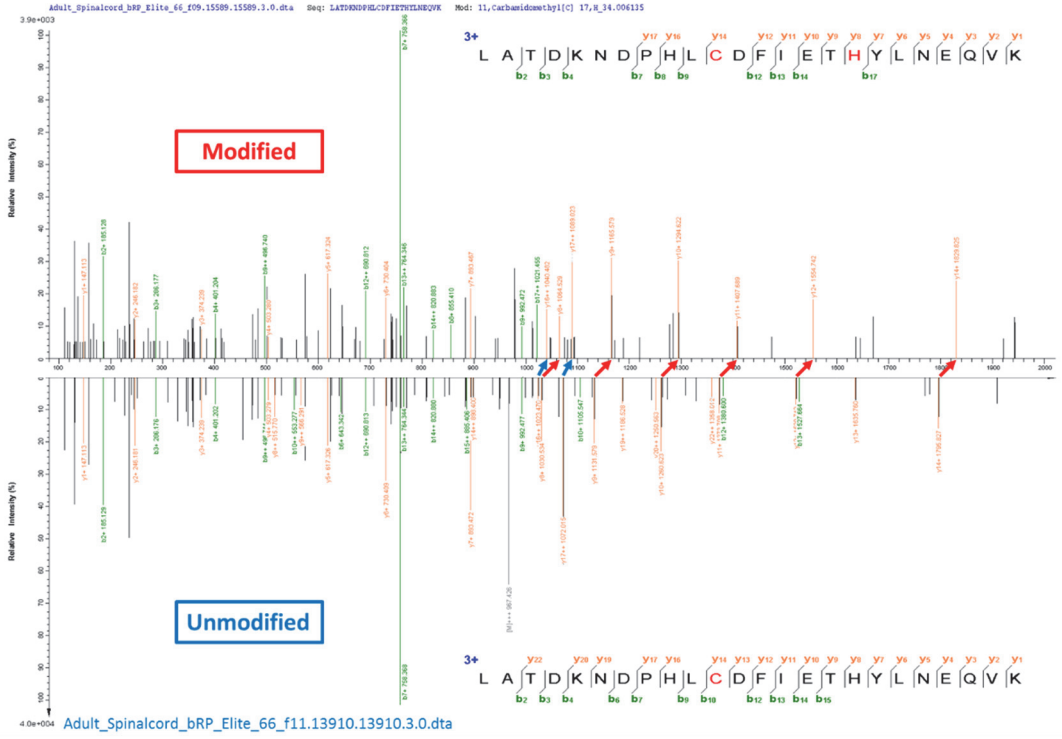
B-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f03.11402.11402.2.0.dta

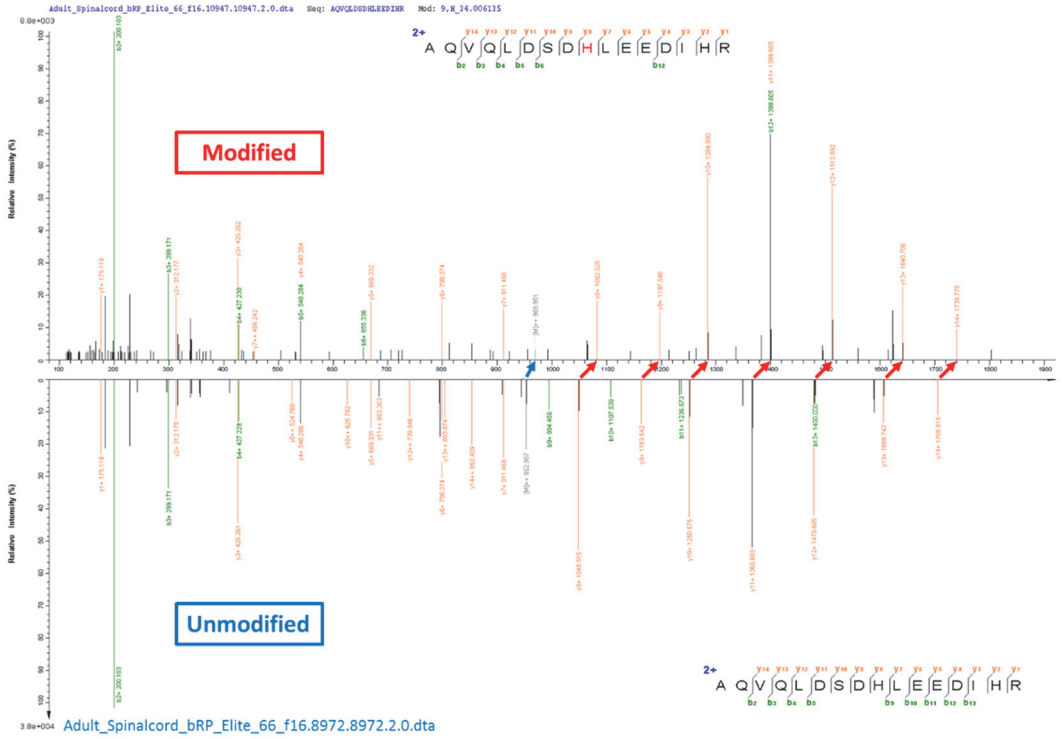
B-2


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f03.11402.11402.2.0.dta

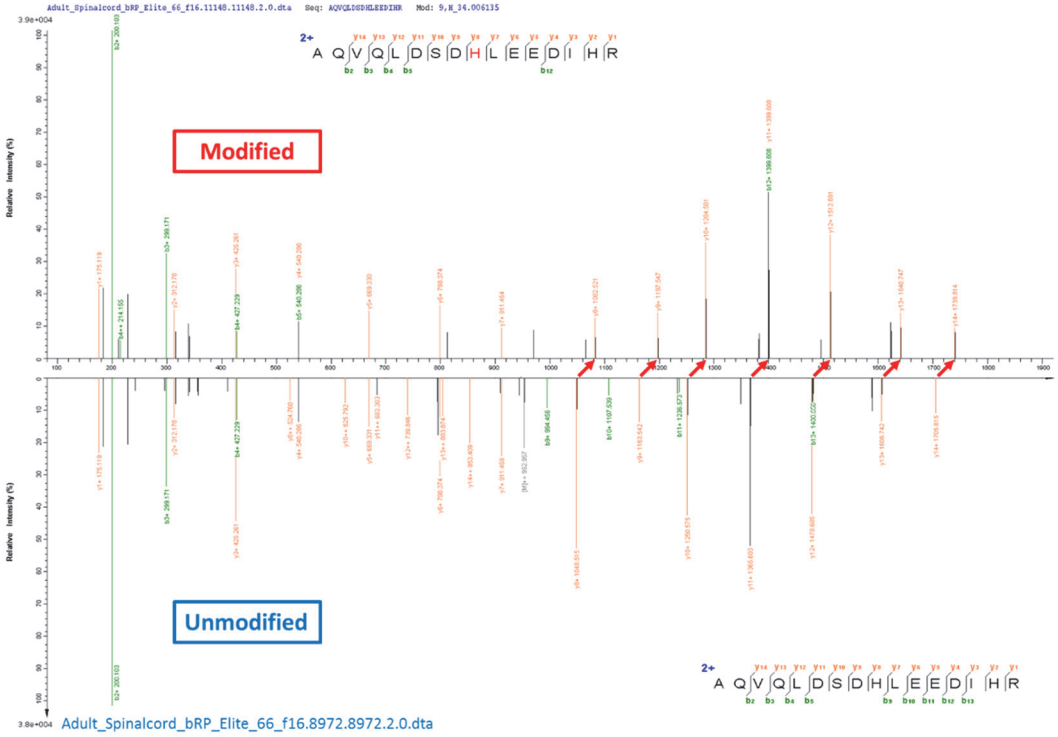
B-3


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f11.13910.13910.3.0.dta

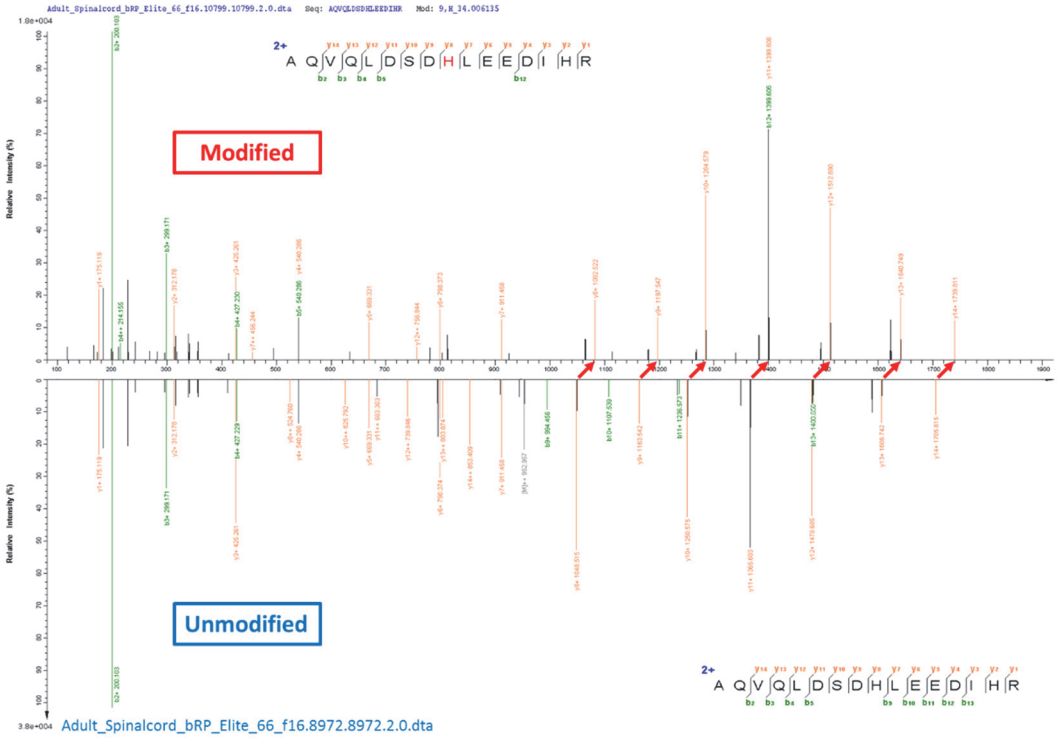
C-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f16.8972.8972.2.0.dta

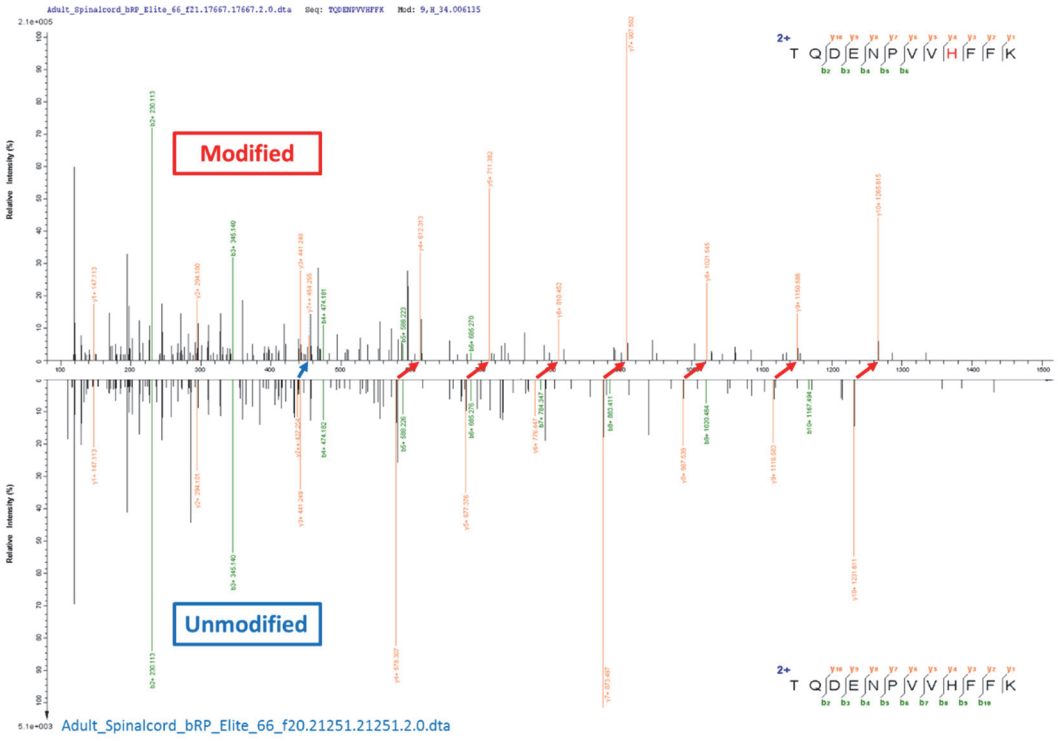
D-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f16.8972.8972.2.0.dta

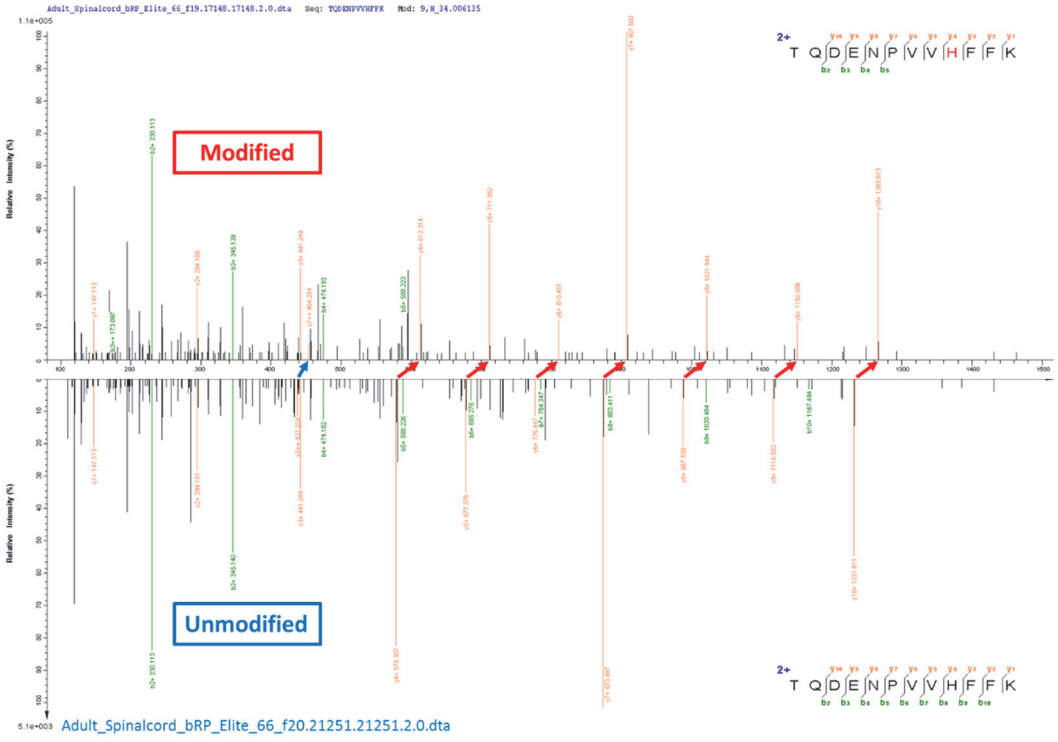
D-2


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f16.8972.8972.2.0.dta

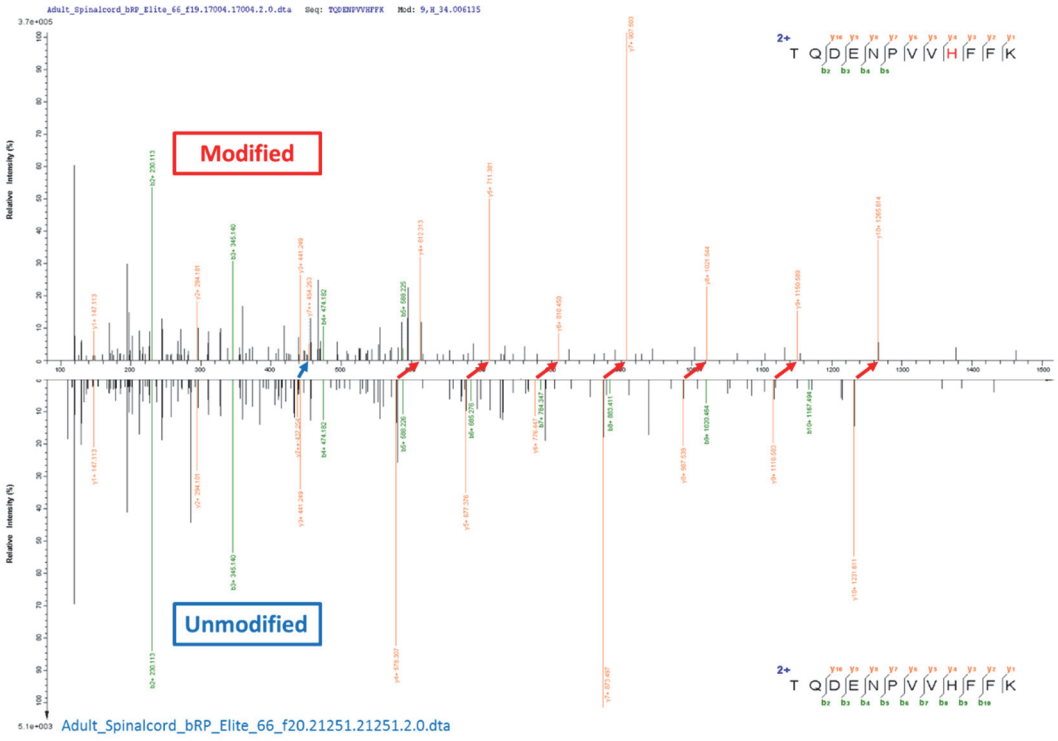
D-3


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.21251.21251.2.0.dta

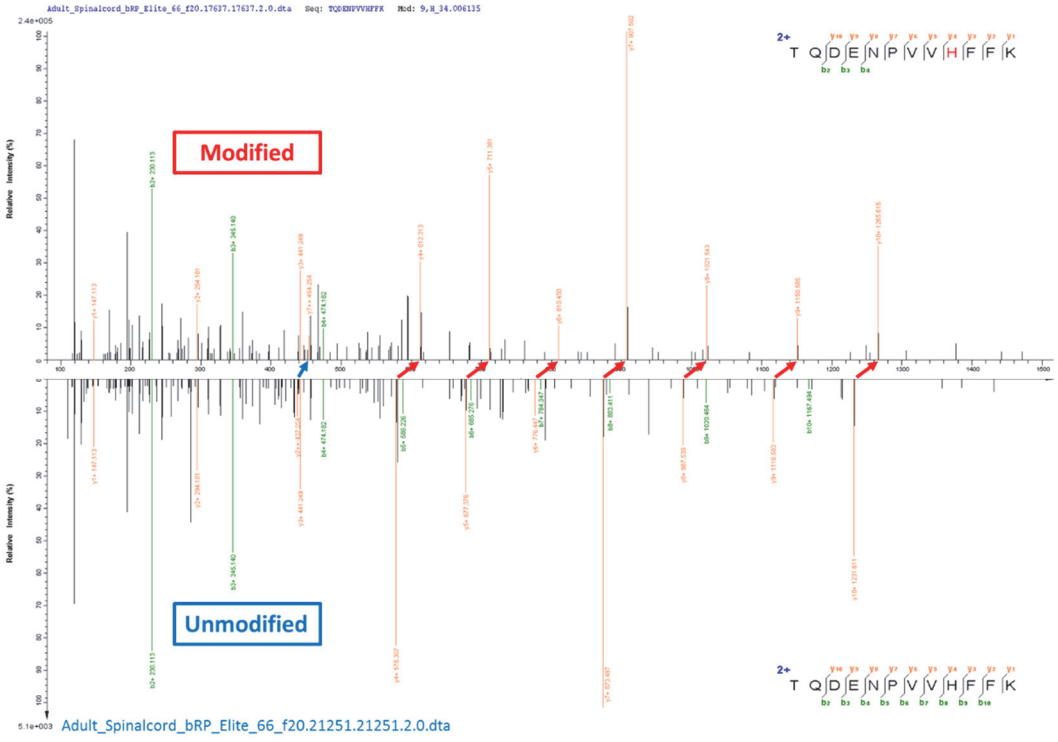
E-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.21251.21251.2.0.dta

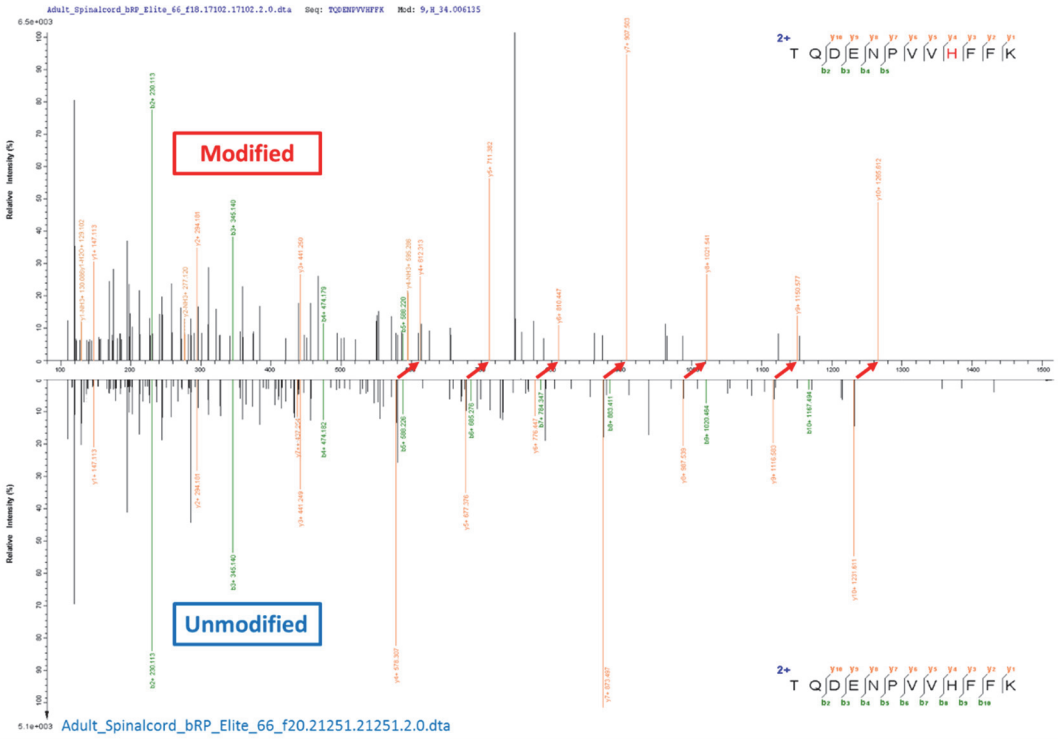
E-2


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.21251.21251.2.0.dta

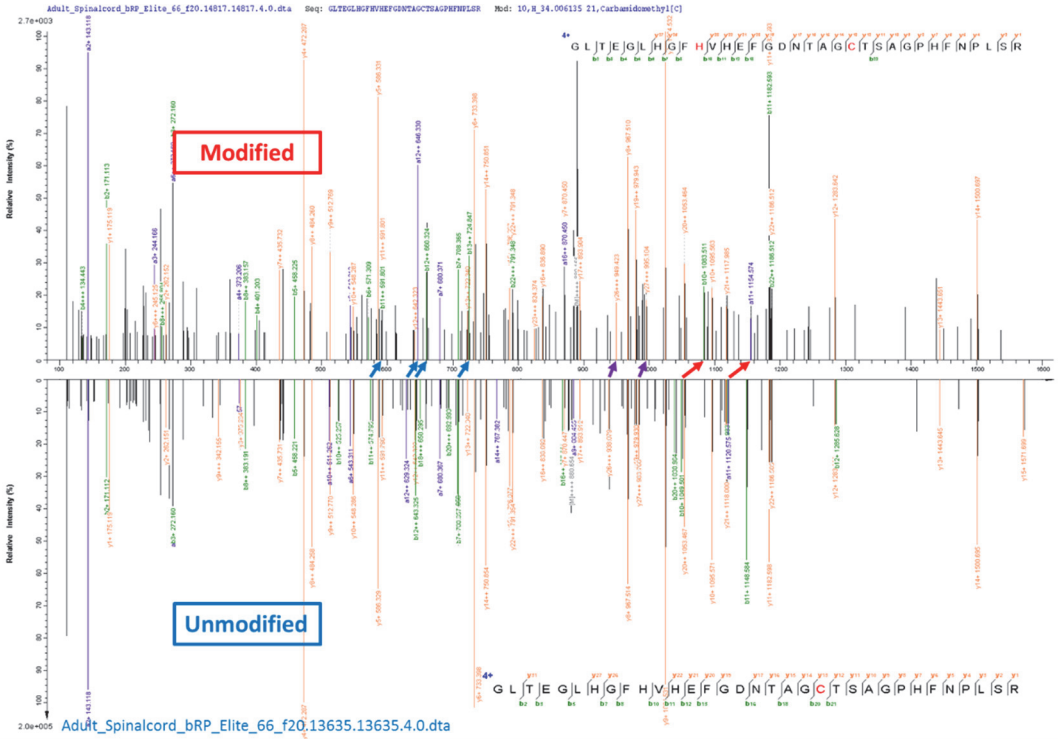
E-3


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.21251.21251.2.0.dta

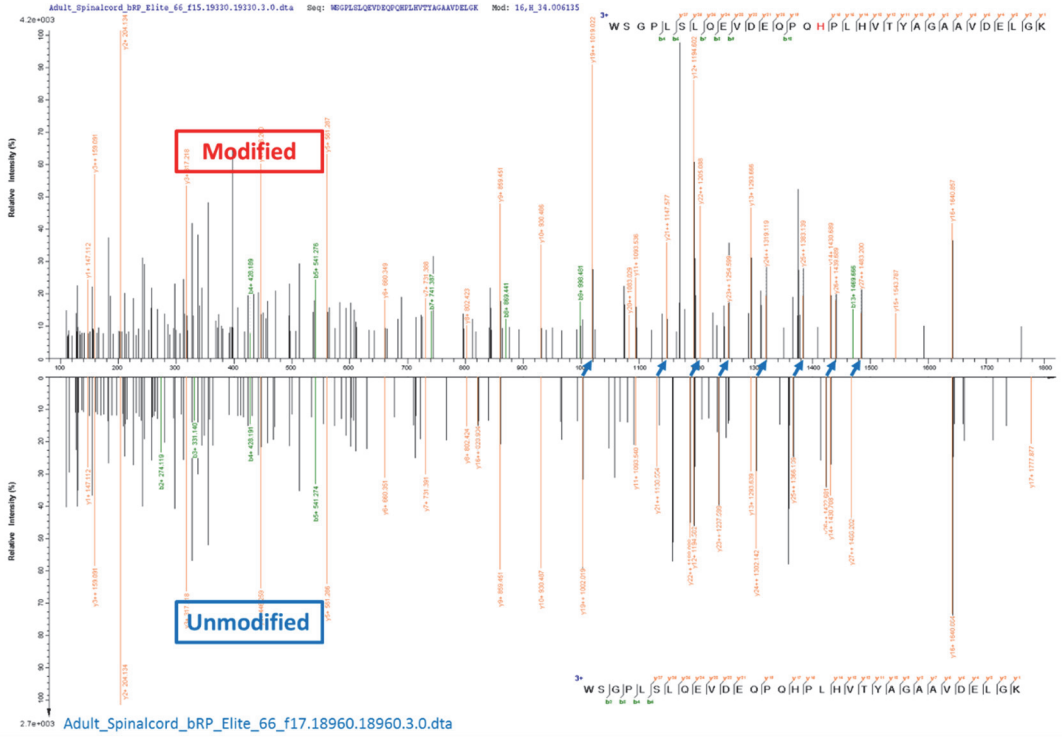
E-4


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.21251.21251.2.0.dta

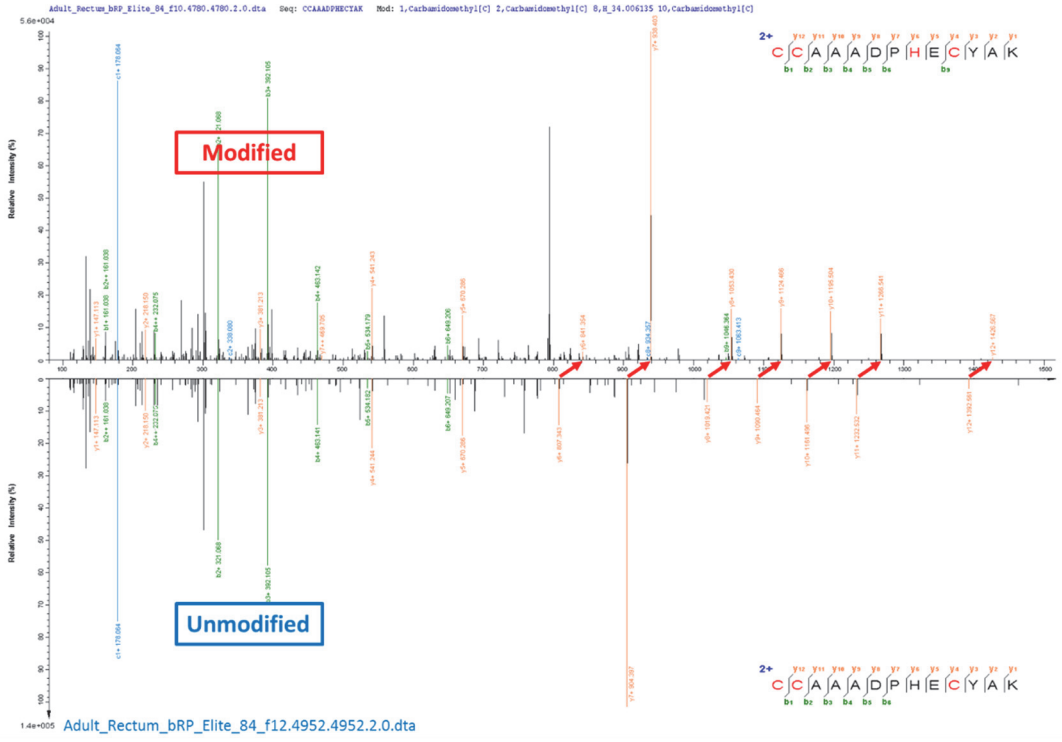
E-5


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f20.13635.13635.4.0.dta

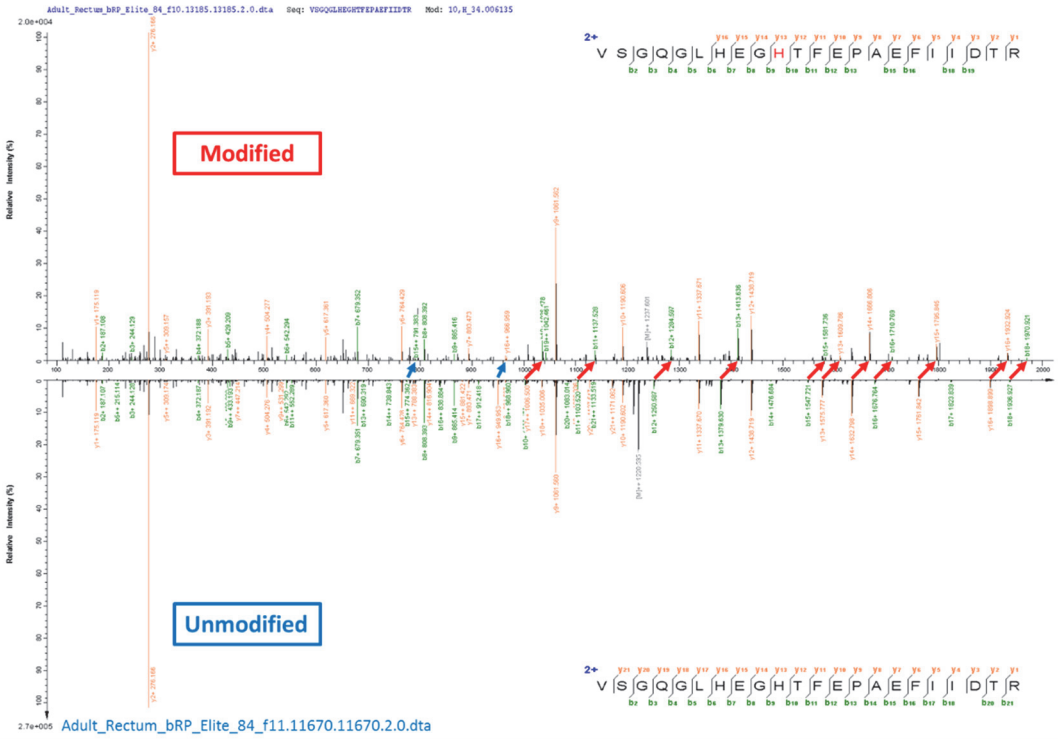
F-1


Modified
Unmodified
Adult_Spinalcord_bRP_Elite_66_f17.18960.18960.3.0.dta

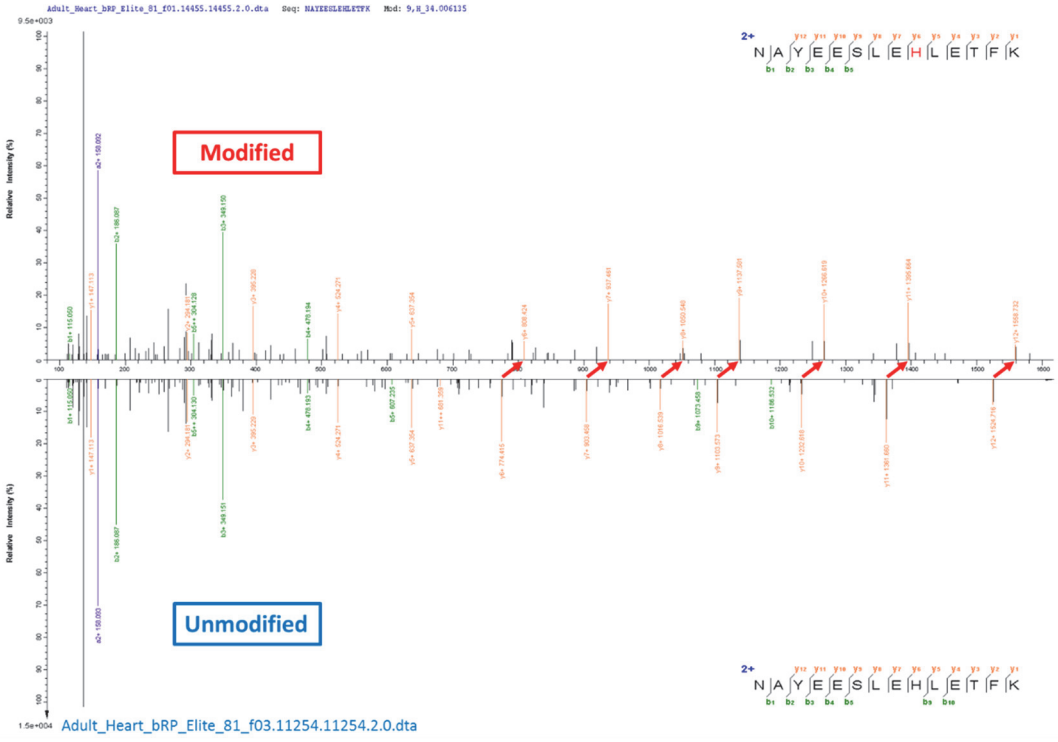
G-1


Modified
Unmodified
Adult_Rectum_bRP_Elite_84_f12.4952.4952.2.0.dta

H-1



I-1



J-1

Supplementary Fig. 15. Twenty representative PSMs with mass shift of 34.006135 Da localized to His. MS/MS spectra of modified (top) and unmodified (bottom) forms of the peptide are shown and compared. Red, blue and purple arrows between two spectra indicate the shift of fragment ion

peaks with one, two and three charges, respectively. The indexes of A-J indicate different peptide sequences.

A (3): TQYEAMASSNMHEAEEWYR

B (3): LEQEHLLEDIAHVR

C (1): LATDKNDPHLCDFIETHYLNEQVK

D (3): AQVQLDSDHLEEDIHR

E (5): LEQEHLLEDIAHVR

F (1): GLTEGLHGFHVHEFGDNTAGCTSAGPHFNPLSR

G (1): WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK

H (1): CCAAADPHECYAK

I (1): VSGQGLHEGHTFEPAEFIIDTR

J (1): NAYEESLEHLETFK

Supplementary Table

Supplementary Table 1. The experimental design of simulated data set.

Supplementary Table 2. The List of modified peptides spiked into HeLa and *E. coli* samples. Here the “ph”, “me”, “ac” and “pr” represent phosphorylation, methylation, acetylation and propionyl, respectively.

Supplementary Table 3. The comparison results between original Ascore and our extended Ascore (see EXCEL file for details).

Supplementary Table 4. The performance of global, separate and transfer FDR approaches at 1% FDR for MODa open search results of the simulated data set. “Sub-Correct” means that the identified peptide is part of the real peptide or in the contrary.

Supplementary Table 5. The annotated modification types and their corresponding numbers given by PTMiner for the draft map of human proteome (see EXCEL file for details).

Supplementary Table 6. Fully annotated modifications with >1000 PSMs in the data of draft map of human proteome.

Supplementary Table 7. Partially annotated modifications with >1000 PSMs in the data of draft map of human proteome.

Supplementary Table 8. Unannotated modifications with >1000 PSMs in the data of draft map of human proteome.

Supplementary Table 9. The list of PSMs with SAVs registered in Uniprot database from the draft map of human proteome.

Supplementary Table 10. Some examples of mass shifts explained as in-source fragmentation, non-specific digestion or missed cleavage events.

Supplementary Table 1. The experimental design of simulated data set.

PTM	Subset	Mass	Number of spectra
Phospho[S]	$S_{phospho,S}$	79.966331	25
Phospho[T]	$S_{phospho,T}$	70.966331	25
Di-methyl[K]	$S_{dimethyl,K}$	28.031300	500
Methyl[K]	$S_{methyl,K}$	14.015650	1,000
Tri-methyl[K]	$S_{trimethyl,K}$	42.046950	5,000
Acetyl[peptide N-term]	$S_{acetyl,N-term}$	42.010565	5,000
Oxidation[W]	$S_{oxidation,W}$	15.994915	10,000
Deamidation[N]	$S_{deamidation,N}$	0.984016	15,000
Oxidation[M]	$S_{oxidation,M}$	15.994915	20,000
Contaminated	$S_{contaminated}$	0	100,000
Unmodified	$S_{unmodified}$	0	800,000
Total	-	-	956,550

Supplementary Table 2. The List of modified peptides spiked into HeLa and *E. coli* samples. Here the "ph", "me", "ac" and "pr" represent phosphorylation, methylation, acetylation and propionyl, respectively.

Uniprot Accession	Gene Name	Species	Sequence
P62805	HIST1H4A	Homo	DNIQGIT(ph)KPAIR
P0A870	talB	*E. coli*	EYAPAED(me)PGVVSVSEIYQYYK
P0A7K2	rplL	*E. coli*	GATGLGLKE(me)AK
P62807	HIST1H2BC[a]	Homo	KESK(ac)YSVYVYK
A4FNV9	sace_6566	*S. erythraea*	TYK(pr)LYVGGK
A4FBI4	sace_2103	*S. erythraea*	HGGGAFSGK(pr)DPSK

[a] Add one Lys between the 37th and 38th amino acid in its protein sequence.

Supplementary Table 3. The comparison results between original Ascore and our extended Ascore (see EXCEL file for details).

Supplementary Table 4. The performance of global, separate and transfer FDR approaches at 1% FDR for MODa open search results of the simulated data set. “Sub-Correct” means that the identified peptide is part of the real peptide or in the contrary.

	Total Spectra	Real FDR			Identification number		
		Global	Separate	Transfer	Global	Separate	Transfer
Unmodified	800,000	0.08%	1.01%	1.01%	338,663	610,533	610,533
Oxidation	30,000	0.28%	1.08%	1.00%	9,547	15,069	13,888
Deamidation	15,000	0.81%	1.07%	1.33%	4,797	5,255	5,405
Tri-Methyl & Acetyl	10,000	0.52%	1.19%	1.19%	3,114	3,949	3,940
Methyl	1,000	12.39%	0.00%	0.00%	323	75	1
Di-Methyl	500	11.18%	4.26%	0.00%	164	47	0
Phospho	50	23.81%	0.00%	0.00%	22	8	0
Sub-Correct	-	0.00%	0.00%	0.00%	255	231	164
Other Modifications	-	100%	100%	100%	3,185	628	21
Sum	856,550	-	-	-	360,070	635,795	633,952

Supplementary Table 5. The annotated modification types and their corresponding numbers given by PTMiner for the draft map of human proteome (see EXCEL file for details).

Supplementary Table 6. Fully annotated modifications with >1,000 PSMs in the data of draft map of human proteome.

Modification	Count	Modification	Count
Acetyl (Any N-term)	27,911	Dethiomethyl (M)	7,238
Acetyl (Protein N-term)	38,841	Dicarbamidomethyl (Any N-term)	2,290
Carbamidomethyl (Any N-term)	142,668	Didehydro (T)	1,057
Carbamidomethyl (D)	3,158	Didehydro (Y)	1,215
Carbamidomethyl (E)	2,931	Dioxidation (M)	5,117
Carbamidomethyl (H)	4,603	Dioxidation (W)	10,598
Carbamidomethyl (K)	2,255	Formyl (Any N-term)	186,463
Carbamyl (Any N-term)	180,856	Formyl (K)	1,484
Carbamyl (K)	12,672	Formyl (S)	8,410
Carbamyl (M)	14,002	Formyl (T)	4,199
Carboxymethyl (Any N-term)	5,141	Gln->pyro-Glu (Any N-term Q)	3,937
Deamidated (N)	101,574	Glu->pyro-Glu (Any N-term E)	1,059
Deamidated (Q)	9,433	Methyl (E)	4,181
Deamidated (R)	2,856	Oxidation (M)	482,046
Dehydrated (D)	1,029	Oxidation (P)	8,885
Dehydrated (T)	1,368	Oxidation (W)	3,000
Delta:H(2)C(2) (Any N-term)	75,208	Succinyl (Any N-term)	3,131

Supplementary Table 7. Partially annotated modifications with >1000 PSMs in the data of draft map of human proteome.

Modification Mass	Count	Modification Mass	Count	Modification Mass	Count
-1.007825	19,362	85.031634	3,419	40.006148	1,693
12	14,247	12.017759	3,171	-9.032697	1,450
-2.01565	11,209	-17.026549	3,131	-2.981907	1,387
-0.984016	10,787	-1.979265	2,622	73.028954	1,282
-1.997892	10,459	39.994915	2,360	-18.010565	1,149
-1.031634	7,927	44.008456	2,241	184.07961	1,119
43.989829	6,444	59.019355	2,178	218.167065	1,072

Supplementary Table 8. Unannotated modifications with >1000 PSMs in the data of draft map of human proteome.

Mass bin in integer	Count	Average mass	Standard deviation
-112	1,104	-112.099	0.009881
-91	2,836	-91.0091	0.010413
-85	1,433	-85.0867	0.02114
-3	5,850	-3.00997	0.021254
-2	1,772	-2.01609	0.073557
-1	7,457	-0.99059	0.085496
1	8,250	1.022004	0.12328
2	2,160	1.990707	0.125942
24	2,090	23.99993	0.017695
28	1,192	27.98817	0.059895
55	1,010	55.00993	0.02125
57	1,468	57.00441	0.074891
59	1,256	58.99774	0.030613
83	1,548	83.03704	0.015139
85	1,978	85.01499	0.024082
170	1,196	170.0991	0.028478
171	3,275	171.0969	0.017873
173	1,826	173.0536	0.008608
184	1,369	184.0986	0.014632
185	1,463	185.1144	0.017837
189	2,263	189.0471	0.005507
199	1,451	199.1114	0.015265
239	3,688	239.1263	0.025904
241	5,889	241.166	0.031565
242	1,214	242.1265	0.045652
243	1,221	243.1184	0.0443
257	1,095	257.1283	0.033297
271	1,070	271.1037	0.042487

Supplementary Table 9. The list of PSMs with SAVs registered in Uniprot database from the draft map of human proteome (see EXCEL file for details).

Supplementary Table 10. Some examples of mass shifts explained as in-source fragmentation, non-specific digestion or missed cleavage events.

Mass shift in integer	Count	Explanation	Average value of mass shifts	Theoretical value	Mass difference
-112	534	-K+Oxidation (M)	112.099167	112.100045	0.000878
-85	1574	-K+Carbamyl (Any N-term)	-85.086992	-85.089146	0.002154
170	634	K+Acetyl (Any N-term)	170.105931	170.105525	0.000406
171	2794	K+Carbamyl (Any N-term)	171.102913	171.100774	0.002139
173	1788	M+Acetyl (Any N-term)	173.053692	173.051055	0.002637
184	2169	R+Formyl (Any N-term)	184.097198	184.096025	0.001173
185	666	K+Carbamidomethyl (Any N-term)	185.117561	185.116424	0.001137
189	2230	M+Carboxymethyl (Any N-term)	189.046993	189.045969	0.001024
199	243	AK	199.131882	199.132070	0.000188
199	1018	R+Carbamyl (Any N-term)	199.108953	199.106924	0.002029
239	3156	K+Nmethylmaleimide (K)	239.128729	239.126988	0.001741
241	4528	I/L+K	241.178586	241.179020	0.000434
242	476	NK	242.136173	242.137890	0.001717
243	659	DK	243.123095	243.121900	0.001195
257	195	SK+Acetyl (Any N-term)	257.141075	257.137555	0.003520
257	422	EK	257.137997	257.137550	0.000447
271	626	CK+Pyro-carbamidomethyl (Any N-term C)	271.101256	271.099065	0.002191

Reference

1. Fu Y, Qian X: **Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry**. *Mol Cell Proteomics* 2014, **13**(5):1359-1368.
2. Fu Y: **Bayesian false discovery rates for post-translational modification proteomics**. *Statistics and Its Interface* 2012, **5**(1):47-59.
3. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al*: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.