Systematic Evaluation of Protein Sequence Filtering Algorithms for Proteoform Identification Using Top-Down Mass Spectrometry (Supplementary Material)

Qiang Kou[1], Si Wu[2], and Xiaowen Liu[1,3,*]

[1]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis
[2]Department of Chemistry and Biochemistry, University of Oklahoma
[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine

1 Software tools for top-down mass spectrometry analysis

Table S1: Software tools for proteoform identification using top-down mass spectrometry

Software	Website	Reference
ProSightPC	`http://proteinaceous.net/product/prosightpc-4-0/`	[1]
MS-TopDown	`http://proteomics.ucsd.edu/software-tools/`	[2]
PIITA	-	[3]
Mascot Top Down	`http://www.matrixscience.com/`	[4]
BUPID Top-Down	`http://www.bumc.bu.edu/cardiovascularproteomics` `/cpctools/bupid-top-down/`	[5]
MS-Align+	`http://bix.ucsd.edu/projects/msalign/`	[6]
Byonic	`http://www.proteinmetrics.com/products/byonic/`	[7]
MS-Align-E	`http://proteomics.informatics.iupui.edu/software/msaligne/`	[8]
ProteinGoggle	`http://proteingoggle.tongji.edu.cn/`	[9, 10]
ProSight Lite	`http://prosightlite.northwestern.edu/`	[11]
Proteoform Suite	`https://github.com/smith-chem-wisc/ProteoformSuite`	[12]
TopPIC	`http://proteomics.informatics.iupui.edu/software/toppic/`	[13]
pTop	`http://pfind.ict.ac.cn/software/pTop/index.html`	[14]
TopMG	`http://proteomics.informatics.iupui.edu/software/topmg/`	[15]
MASH Suite Pro	`http://crb.wisc.edu/yinglab/software.html`	[16, 17]
MSPathFinder	`https://omics.pnl.gov/software/mspathfinder`	[18]

Table S2: Software tools for top-down spectral deconvolution

Software	Website	Reference
THRASH	`http://proteinaceous.net/product/prosightpc-4-0/`	[19]
Xtract	`http://proteinaceous.net/product/prosightpc-4-0/`	[20]
DeconMSn	`https://omics.pnl.gov/software/deconmsn`	[21]
YADA	`http://pcarvalho.com/patternlab/downloads/windows/yada/`	[22]
DeconTools (Decon2LS)	`https://omics.pnl.gov/software/decontools-decon2ls`	[23]
MS-Deconv	`http://bix.ucsd.edu/projects/msdeconv/`	[24]
MS-Deconv+/TopFD[1]	`https://github.com/toppic-suite/toppic-suite`	[25]
pParse	`http://pfind.net/software/pTop/index.html`	[26]
UniDec	`http://unidec.chem.ox.ac.uk`	[27]
ProMex	`https://omics.pnl.gov/software/mspathfinder`	[18]

The ASF-RESTRICT Algorithm

Input: A deconvoluted top-down MS/MS spectrum S, a set Ω of f variable PTMs, a number k of intervals, parameters h and t, and a protein database D.

Output: Top t candidate protein sequences in D for the query spectrum S.

```
1. Set the protein set $\Phi$ as an empty set, and compute $k$ intervals as well as their $k$ centers in $S$.
2. For each set of $h$ masses selected from the $k$ centers with replacement do
3.     For each set of $h$ PTMs selected from $\Omega$ with replacement do
4.         Generate an approximate spectrum $S'$ using the $h$ selected masses and the $h$ selected PTMs.
5.         Use the UPF-RESTRICT algorithm to search $S'$ against $D$ to find top $t$ candidate proteins as well as their similarity scores, and add them to $\Phi$.
6. Report $t$ top scoring protein sequences from $\Phi$.
```

Figure S1: The ASF-RESTRICT algorithm for protein sequence filtration using top-down MS/MS spectra.

[1] The MS-Deconv+ algorithm has been implemented and integrated into TopFD.

```
Algorithm
Input: A deconvoluted top-down MS/MS spectrum $S$ with a precursor mass
       $M$ and peaks $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$, where $a_i$ is the $i$th mass and
       $b_i$ is the intensity of $a_i$; $h$ guessed prefix residue masses $c_1 \leq c_2 \leq \ldots c_h$;
       and $h$ guessed PTMs and their corresponding mass shifts $\delta_1, \delta_2, \ldots, \delta_h$.
Output: An approximate spectrum $S'$.
Set $q_0 = 0$, $q_{h+1} = M$, and $q_k = c_k$ for $1 \leq k \leq h$.
For $i = 1$ to $n$ do
    Find two values $q_j$ and $q_{j+1}$ such that $q_j \leq a_i < q_{j+1}$.
    $a'_i = a_i - \sum_{k=1}^{j} \delta_k$.
    If $a'_i > 0$ then add $(a'_i, b_i)$ as a peak to $S'$.
Set the precursor mass of $S'$ as $M - \sum_{k=1}^{h} \delta_k$ and output $S'$.
```

Figure S2: An algorithm for generating an approximate spectrum from a query top-down deconvoluted MS/MS spectrum and a list of guessed prefix residue masses and variable PTMs.

Table S3: Parameter settings of TopPIC in the analysis of the EC data set

Parameter	**Value**
Fragmentation method	FILE
Fixed modifications	None
N-terminal forms of proteins	NONE, NME, NME+ACETYLATION
Using a decoy database	Yes
Error tolerance	15 ppm
Maximum number of unexpected modifications (unknown mass shifts) in a PrSM	0
Cutoff type	Spectrum-level FDR
Cutoff value	0.01
Using the generating function approach to compute p-values and E-values	No
Number of combined spectra	1
Common modifications for characterization of unknown mass shifts	None
E-value computation	Lookup table

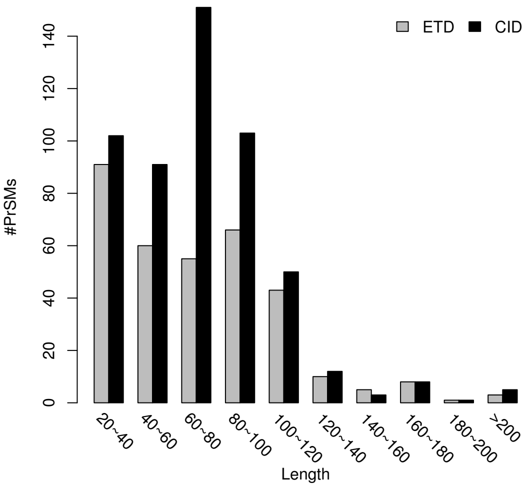


Figure S3: The histogram of the proteoform lengths of the 874 PrSMs identified by TopPIC from the EC data set

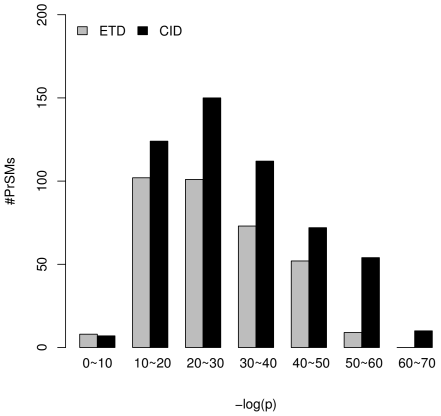


Figure S4: The histogram of the conditional spectral probabilities p of the 874 PrSMs identified by TopPIC from the EC data set

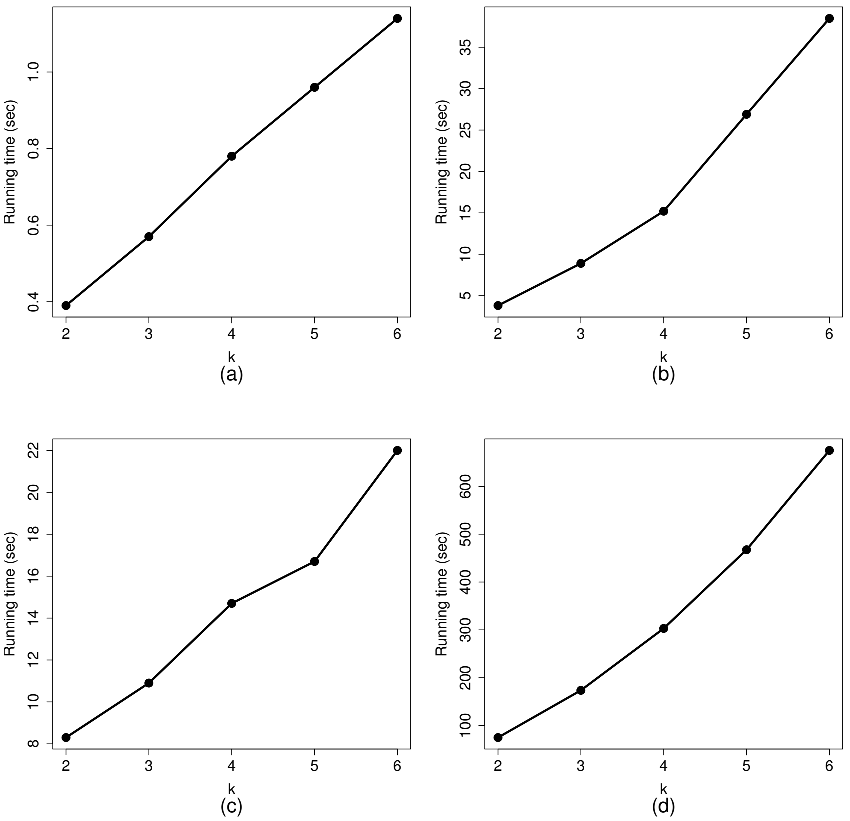


Figure S5: The average running times (per spectrum) of the ASF algorithms with various settings of k and h on the simulated PrSMs with 5 PTMs: (a) ASF-RESTRICT with $h = 1$; (b) ASF-RESTRICT with $h = 2$; (c) ASF-DIAGONAL with $h = 1$; (d) ASF-DIAGONAL with $h = 2$.

2 Tag-based filtering algorithms

A sequence tag is a short amino acid sequence extracted from an MS/MS spectrum. Most tag extraction methods are based on spectrum graphs [28]. A spectrum graph is constructed from a deconvoluted MS/MS spectrum using three steps (Fig. S6): (a) A node is added to the spectrum graph for each fragment mass in the spectrum. (b) Two nodes are connected by an edge if the difference between their corresponding masses is similar to (within an error tolerance) the mass of an amino acid residue[2]. The label of the edge is the amino acid. (c) A node is removed from the graph if there are no edges connecting to it. Each path in the spectrum graph corresponds to a sequence tag. A top-down spectrum graph typically consists of several connected components because of many missing peaks.

We describe two sequence tag-based filtering methods, which are used in MS-Align+Tag [29] and MSPathFinder [18], respectively. The first method uses the long tag strategy to obtain sequence tags from a spectrum graph with three steps: (a) A longest sequence tag is selected from each component of

[2] In some tag generation methods, two nodes are connected if their corresponding mass difference is similar to the mass of one or two amino acid residues.

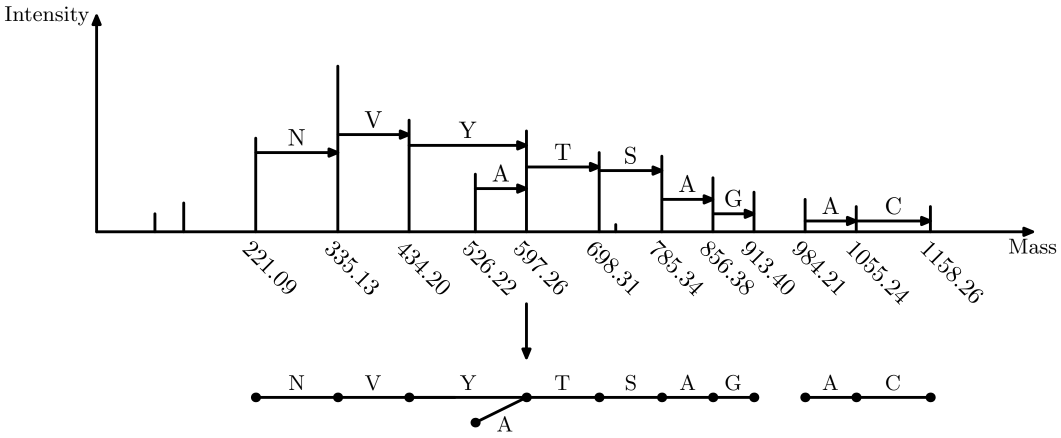


Figure S6: A spectrum graph (bottom) is constructed from a deconvoluted MS/MS spectrum (top). The two left most nodes correspond to masses 221.09 Da and 335.13 Da in the spectrum. These two nodes are connected by an edge because the difference between 221.09 and 335.13 is similar to the mass of an asparagine residue (114.04 Da). The spectrum graph contains two connected components.

the spectrum graph. If a component contains several longest sequence tags with the same length, one of them is arbitrarily selected. (b) The reported sequence tags are filtered to remove those with less than k amino acids ($k = 4$ in the experiments). (c) For each remaining sequence tag, all of its substrings with length k are reported. For example, in Fig. S6, the longest sequence tags NVYTSAG and AC are extracted from the spectrum graph, then the tag AC is filtered out because its length is less than $k = 4$, and finally four length-4 short tags are extracted: NVYT, VYTS, YTSA, and TSAG.

In the second method, we extract from the spectrum graph all sequence tags with a length l between the minimum length l_{min} and the maximum length l_{max}, that is, $l_{min} \leq l \leq l_{max}$. In the experiment, $l_{min} = 5$ and $l_{max} = 8$. First, all tags with length l_{max} are extracted from the spectrum graph and added to a sequence tag set T. For example, when $l_{max} = 6$, two tags NVYTSA and VYTSAG are extracted from the graph in Fig. S6. Next, all tags with length $l_{max} - 1$ are extracted. A length $l_{max} - 1$ tag is added to T if it is not a substring of any tag in T. For example, the length-5 sequence tag NVYTS in Fig. S6 is not added to T because it is a substring of the length-6 sequence tag NVYTSA, and the sequence tag ATSAG is added to T because it is not a substring of any tag in T. Two tags in T may share a substring, but their whole sequences are different. Similarly, we further extract sequence tags with lengths $l_{max} - 2, \ldots, l_{min}$ and add them to T if they are not substrings of tags in T. The two methods are called TAG-LONG (with the long tag strategy) and TAG-VAR (with tags of various lengths), respectively.

Because some sequence tags are extracted from suffix fragment ion series, a reversed tag is generated from each extracted tag. The extracted sequence tags and their reversed tags are searched against a protein database to find a small number of top candidate proteins. Because the lengths of proteins vary significantly from dozens to tens of thousands, we compute similarity scores between sequence tags and protein fragments with similar lengths rather than whole proteins. Protein fragments are generated using a parameter L ($L = 150$ in the experiments). If the length of a protein is no larger than L, the whole protein sequence is a fragment. Otherwise, each length L substring of the protein is a fragment, and the total number of fragments of the protein is $n - L + 1$.

Let T be a set of sequence tags and reversed tags extracted from a spectrum graph. We define a similarity score between a candidate fragment and T. If a sequence tag is a substring of a fragment, we say the sequence tag has a hit in the fragment. The *tag score* between the fragment and T is the number of tags in T that have a hit in the fragment. The *tag score* between a protein and T is the maximum tag score among its fragments. All proteins in the protein database are ranked based on their tag scores and the top t ($t = 20$ in experiments) proteins are reported as filtering results.

Table S5: Parameter settings of the tag-based, UPF-based, and ASF-based filtering algorithms

Parameter	**TAG-LONG**	**TAG-VAR**	**UPF-RESTRICT** **UPF-DIAGONAL**	**ASF-RESTRICT** **ASF-DIAGONAL**
Fixed modifications	None	None	None	None
Error tolerance	15 ppm[3]	15 ppm	15 ppm	15 ppm
# threads			1	1
l	4			
l_{min}		5		
l_{max}		8		
k				3
h				1

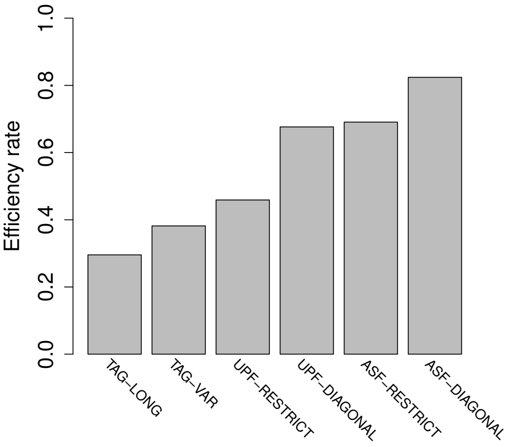


Figure S7: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS and ASF algorithms on the simulated test PrSMs with 5 PTMs.

[3] For the tag-based algorithms, two nodes in a spectrum graph corresponding to two masses $m_1 < m_2$ are connected by an edge if the mass difference $m_2 - m_1$ matches the residue mass of an amino acid within an error tolerance $15 \times 10^{-6} \cdot (m_2 + m_1)/2$.

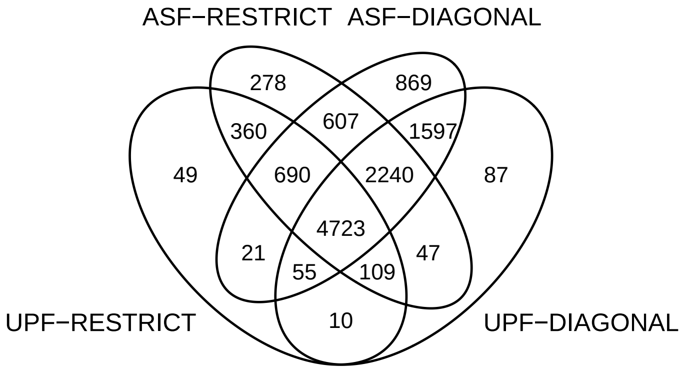


Figure S8: Comparison of the simulated test PrSMs with 5 PTMs that are efficiently filtered by the UPS-RESTRICT, UPS-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms.

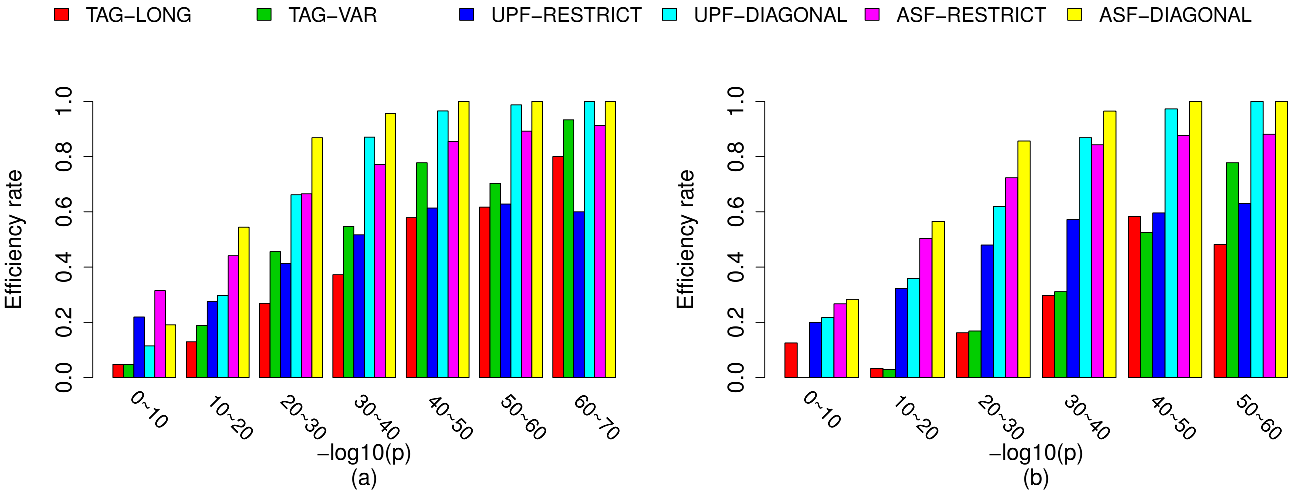


Figure S9: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS and ASF algorithms on PrSMs with various conditional spectral probabilities. The simulated test PrSMs with 5 PTMs are divided into 7 groups based on their conditional spectral probabilities p. (a) CID spectra, (b) ETD spectra.

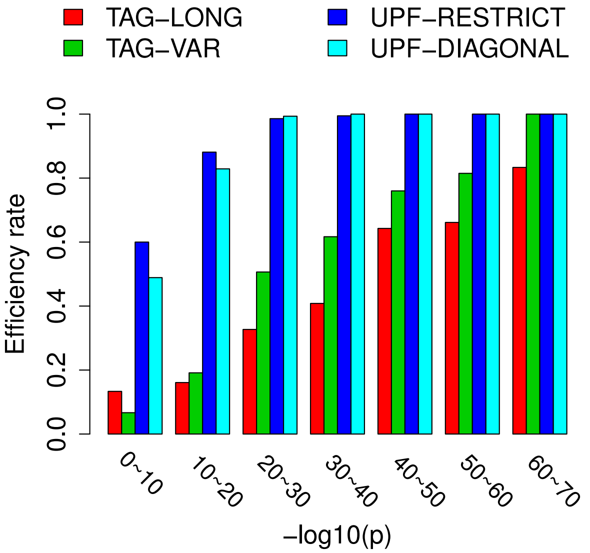


Figure S10: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS-RESTRICT and UPS-DIAGONAL algorithms on the simulated test PrSMs with 1 PTM. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rate for each group is compared.

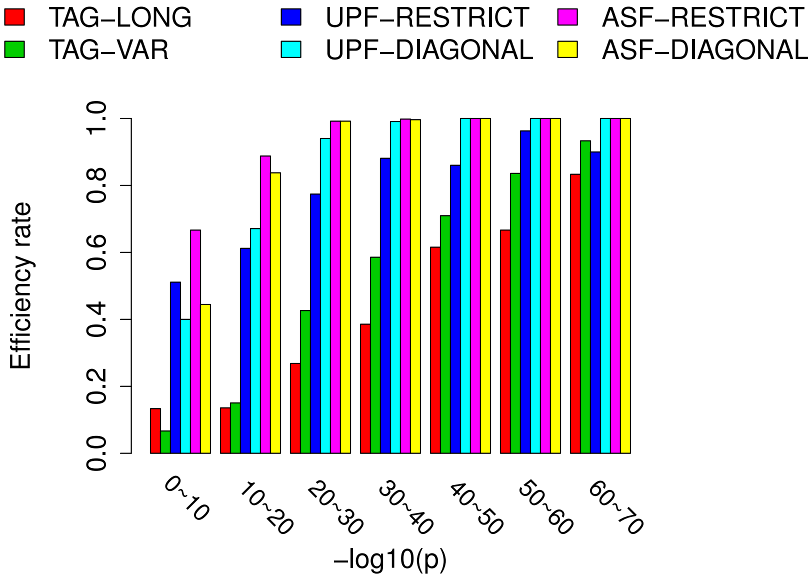


Figure S11: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS and ASF algorithms on the simulated test PrSMs with 2 PTMs. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rate for each group is compared.

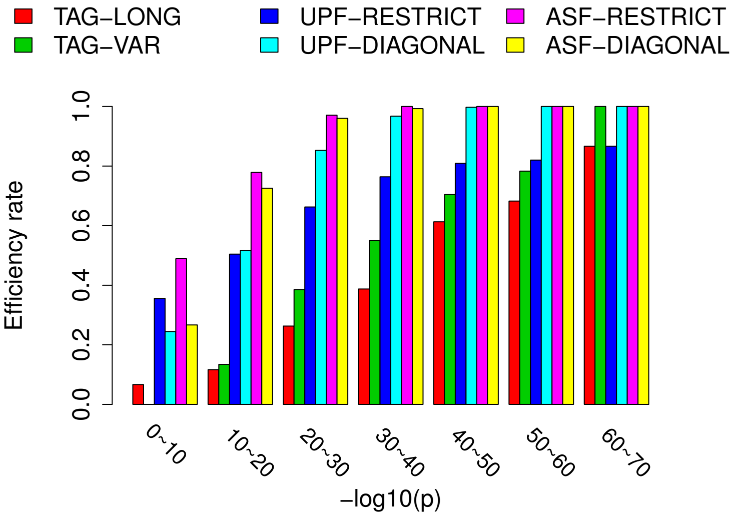


Figure S12: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS and ASF algorithms on the simulated test PrSMs with 3 PTMs. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rate for each group is compared.

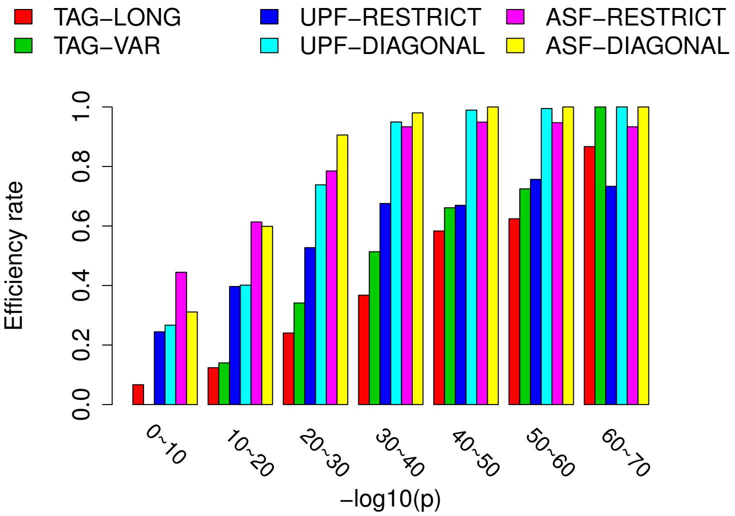


Figure S13: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS and ASF algorithms on the simulated test PrSMs with 4 PTMs. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rate for each group is compared.

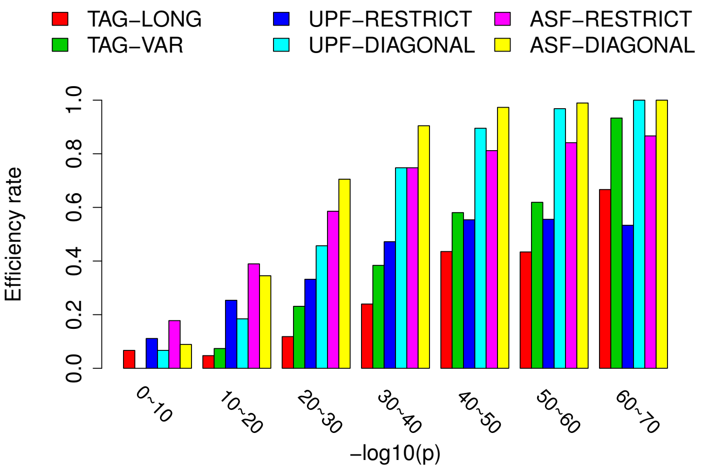


Figure S14: Comparison of the filtering efficiency rates of the TAG-LONG, TAG-VAR, UPS-RESTRICT, UPS-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms on the simulated test PrSMs with 5 PTMs using the EC proteome database concatenated with the human proteome database. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rate for each group is compared.

Table S6: Five variable PTMs used in the identification of proteoforms of the histone H3 and H4 proteins

PTM	Monoisotopic mass shift (Da)	Amino acids
Acetylation	42.01056	R, K
Methylation	14.01565	R, K
Dimethylation	28.03130	R, K
Trimethylation	42.04695	R
Phosphorylation	79.96633	S, T, Y

Table S7: Parameter settings of TopMG in the analysis of the histone data sets

Parameter	**Value**
Fragmentation method	FILE
Fixed modifications	None
N-terminal forms of proteins	NONE, NME, NME+ACETYLATION
Using a decoy database	No
Error tolerance	0.1 Da
Maximum number of unexpected modifications (unknown mass shifts) in a PrSM	0
Number of combined spectra	1
Gap in constructing proteoform graph	40
Maximum number of variable modifications	10

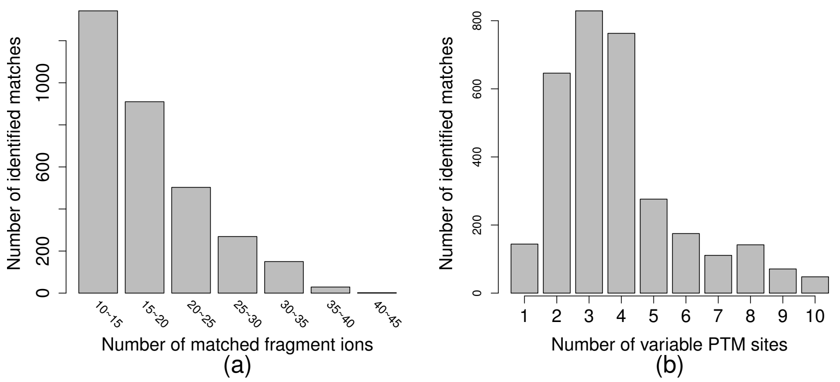


Figure S15: Histograms of the 3 205 PrSMs identified from the histone H3 data set: (a) the number of matched fragment ions, (b) the number of variable PTM sites.

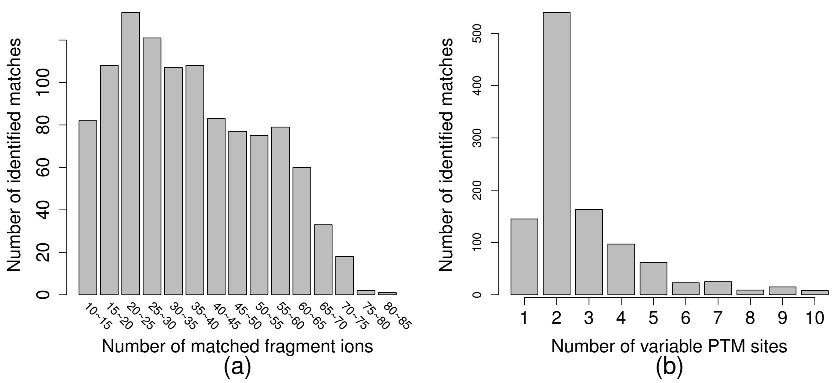


Figure S16: Histograms of the 1 087 PrSMs identified from the histone H4 data set: (a) the number of matched fragment ions, (b) the number of variable PTM sites.

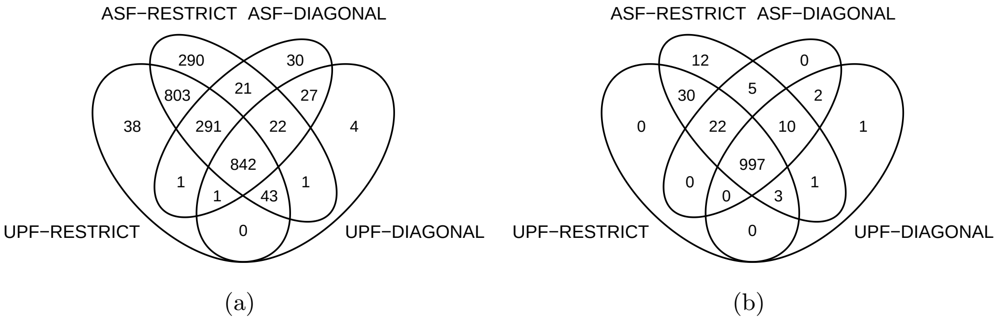


Figure S17: Comparison of the numbers of PrSMs efficiently filtered by the UPF-RESTRICT, UPF-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms: (a) comparison on the 3 205 histone H3 PrSMs; (b) comparison on the 1 087 histone H4 PrSMs.

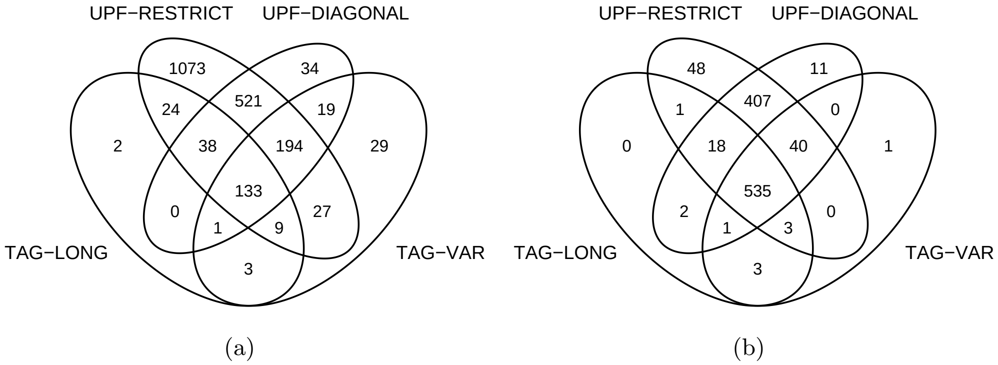


Figure S18: Comparison of the numbers of PrSMs efficiently filtered by the TAG-LONG, TAG-VAR, ASF-RESTRICT and ASF-DIAGONAL algorithms: (a) on the 3 205 histone H3 PrSMs; (b) on the 1 087 histone H4 PrSMs.

Table S10: Parameter settings of TopPIC in the analysis of the xenograft data set

Parameter	**Value**
Number of combined spectra	1
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	None
Maximum number of unexpected modifications	0
Error tolerance	10 ppm
Cutoff type	Proteoform-level FDR
Cutoff value	0.05
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Thread number	12
E-value computation	Lookup table

Table S11: Parameter settings of TopMG in the analysis of the xenograft data set

Parameter	**Value**
Number of combined spectra	1
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	None
Maximum number of unexpected modifications	0
Error tolerance	10 ppm
Cutoff type	Proteoform-level FDR
Cutoff value	0.05
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Thread number	12
E-value computation	Lookup table
Variable PTM	Phosphorylation
Gap in proteoform graph	40

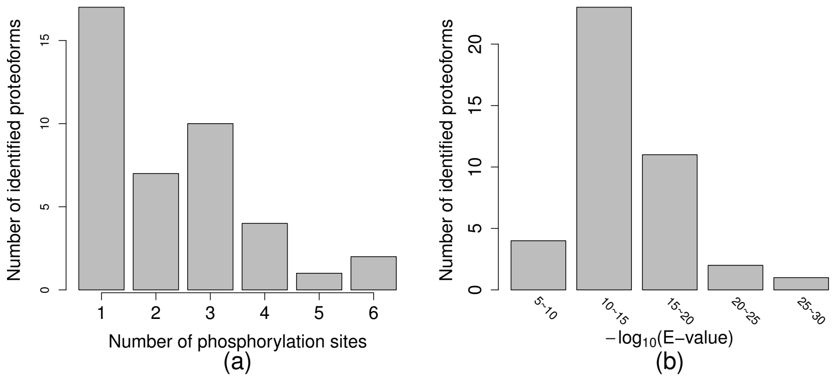


Figure S19: Histograms of the 41 phosphorylated mouse proteoforms identified from the xenograft data set by TopMG: (a) the number of phosphorylation sites, (b) *E*-values.

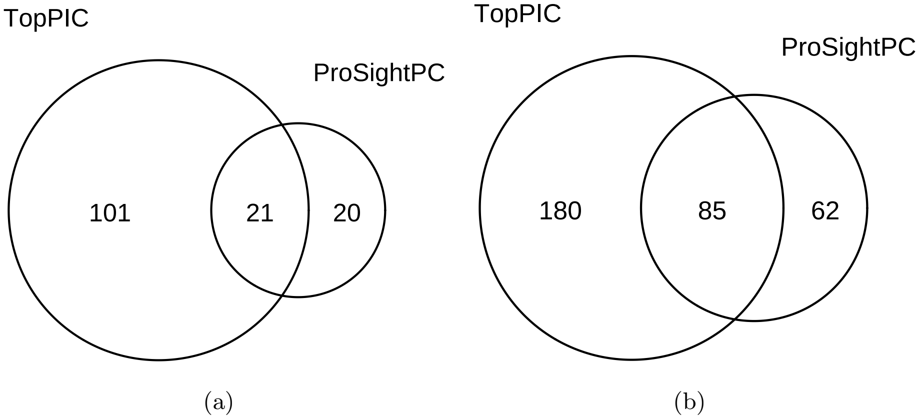


Figure S20: Comparison of the numbers of mouse proteoforms without PTMs (except for terminal truncations and N-terminal acetylation) identified from the xenograft data set by ProSightPC and TopPIC.(a) Mouse proteoforms. (b) Human proteoforms.

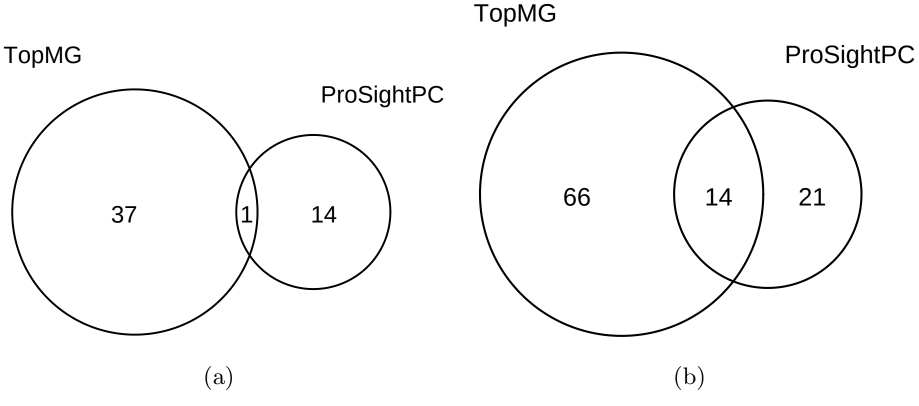
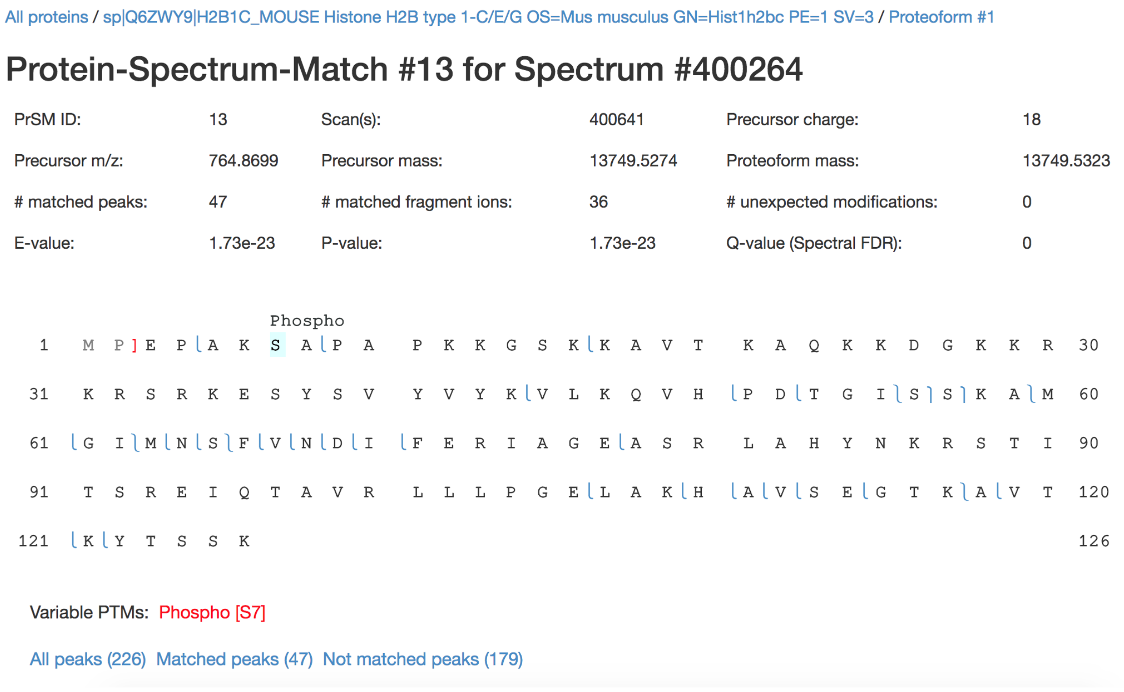


Figure S21: Comparison of the numbers of distinct precursor masses corresponding to the phosphorylated proteoforms identified from the xenograft data set by ProSightPC and TopMG. (a) Mouse proteoforms. (b) Human proteoforms.



All proteins / sp|Q6ZWY9|H2B1C_MOUSE Histone H2B type 1-C/E/G OS=Mus musculus GN=Hist1h2bc PE=1 SV=3 / Proteoform #1

Protein-Spectrum-Match #13 for Spectrum #400264

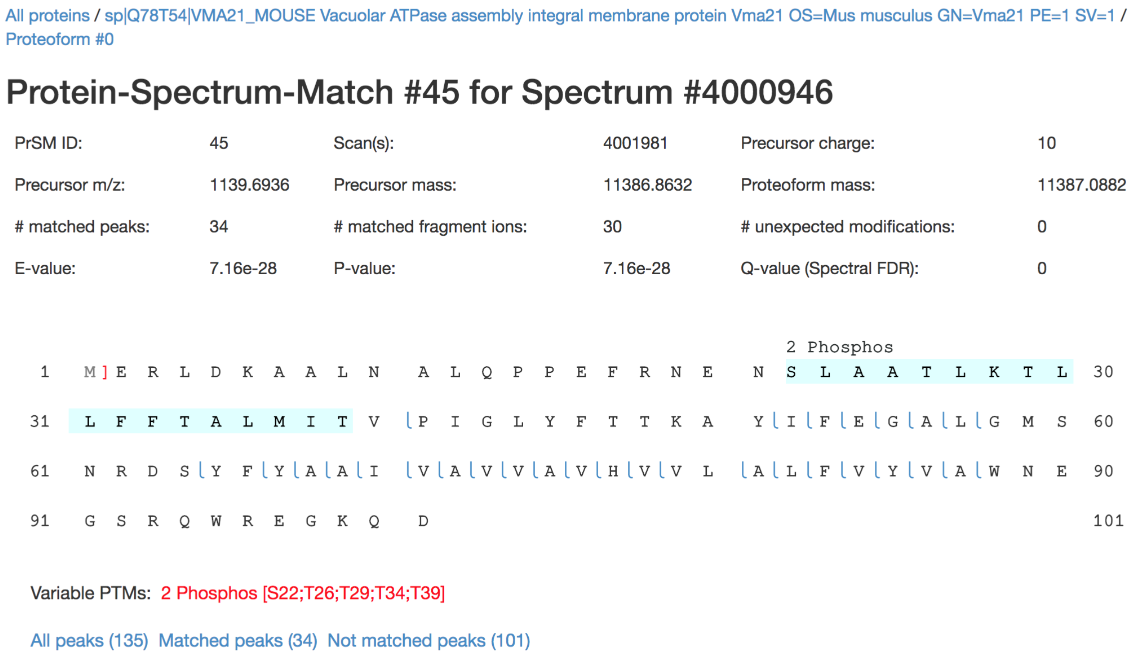
PrSM ID:	13	Scan(s):	400641	Precursor charge:	18
Precursor m/z:	764.8699	Precursor mass:	13749.5274	Proteoform mass:	13749.5323
# matched peaks:	47	# matched fragment ions:	36	# unexpected modifications:	0
E-value:	1.73e-23	P-value:	1.73e-23	Q-value (Spectral FDR):	0

```
                    Phospho
  1   M P]E P⌊A K S A⌊P A   P K K G S K⌊K A V T   K A Q K K D G K K R   30

 31   K R S R K E S Y S V   Y V Y K⌊V L K Q V H  ⌊P D⌊T G I⌉S⌉S⌉K A⌉M   60

 61  ⌊G I⌉M⌊N⌊S⌉F⌊V⌊N⌊D⌊I  ⌊F E R I A G E⌊A S R   L A H Y N K R S T I   90

 91   T S R E I Q T A V R   L L L P G E⌊L A K⌊H  ⌊A⌊V⌊S E⌊G T K⌉A⌊V T  120

121  ⌊K⌊Y T S S K                                                       126
```

Variable PTMs: Phospho [S7]

All peaks (226) Matched peaks (47) Not matched peaks (179)

Figure S22: A phosphorylated proteoform of the mouse histone H2B type 1-C/E/G with one phosphorylation site (UniProt ID: Q6ZWY9) identified by TopMG.



All proteins / sp|Q78T54|VMA21_MOUSE Vacuolar ATPase assembly integral membrane protein Vma21 OS=Mus musculus GN=Vma21 PE=1 SV=1 / Proteoform #0

Protein-Spectrum-Match #45 for Spectrum #4000946

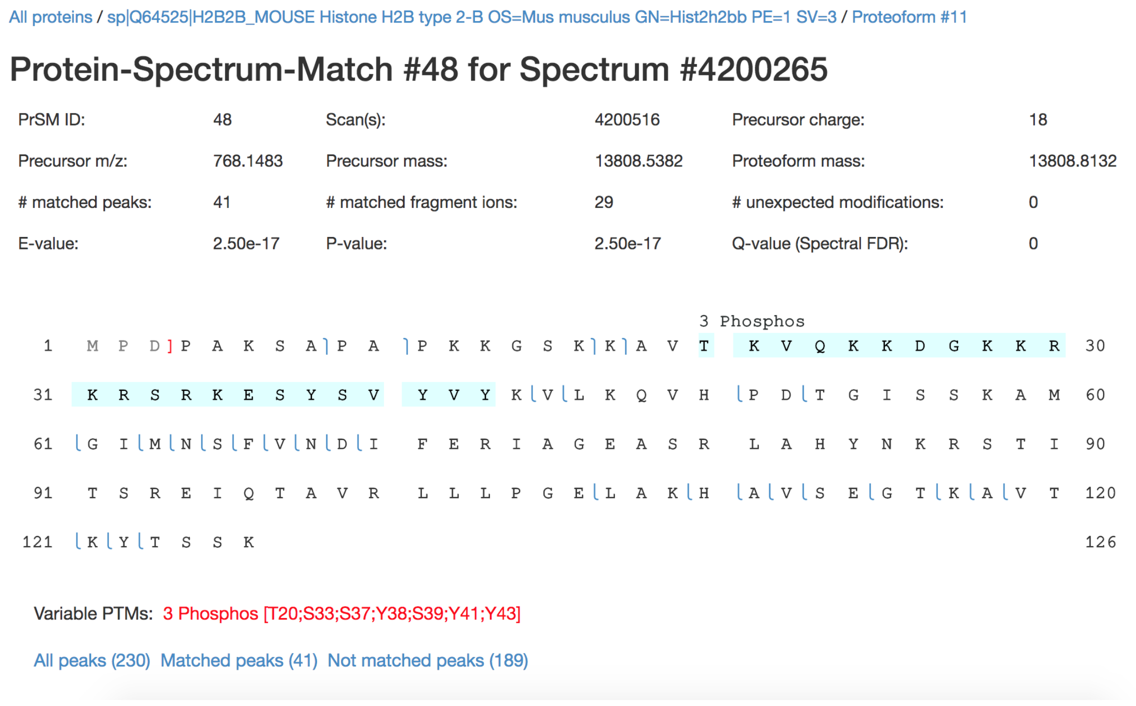
PrSM ID:	45	Scan(s):	4001981	Precursor charge:	10
Precursor m/z:	1139.6936	Precursor mass:	11386.8632	Proteoform mass:	11387.0882
# matched peaks:	34	# matched fragment ions:	30	# unexpected modifications:	0
E-value:	7.16e-28	P-value:	7.16e-28	Q-value (Spectral FDR):	0

```
                                                 2 Phosphos
  1   M]E R L D K A A L N   A L Q P P E F R N E   N S L A A T L K T L   30

 31   L F F T A L M I T V  ⌊P I G L Y F T T K A   Y⌊I⌊F⌊E⌊G⌊A⌊L⌊G M S   60

 61   N R D S⌊Y F⌊Y⌊A⌊A⌊I  ⌊V⌊A⌊V⌊V⌊A⌊V⌊H⌊V⌊V L  ⌊A⌊L⌊F⌊V⌊Y⌊V⌊A⌊W N E   90

 91   G S R Q W R E G K Q   D                                          101
```

Variable PTMs: 2 Phosphos [S22;T26;T29;T34;T39]

All peaks (135) Matched peaks (34) Not matched peaks (101)

Figure S23: A phosphorylated proteoform of the mouse vacuolar ATPase assembly integral membrane protein Vma21 (UniProt ID: Q78T54) with two phosphorylation sites identified by TopMG.



All proteins / sp|Q64525|H2B2B_MOUSE Histone H2B type 2-B OS=Mus musculus GN=Hist2h2bb PE=1 SV=3 / Proteoform #11

Protein-Spectrum-Match #48 for Spectrum #4200265

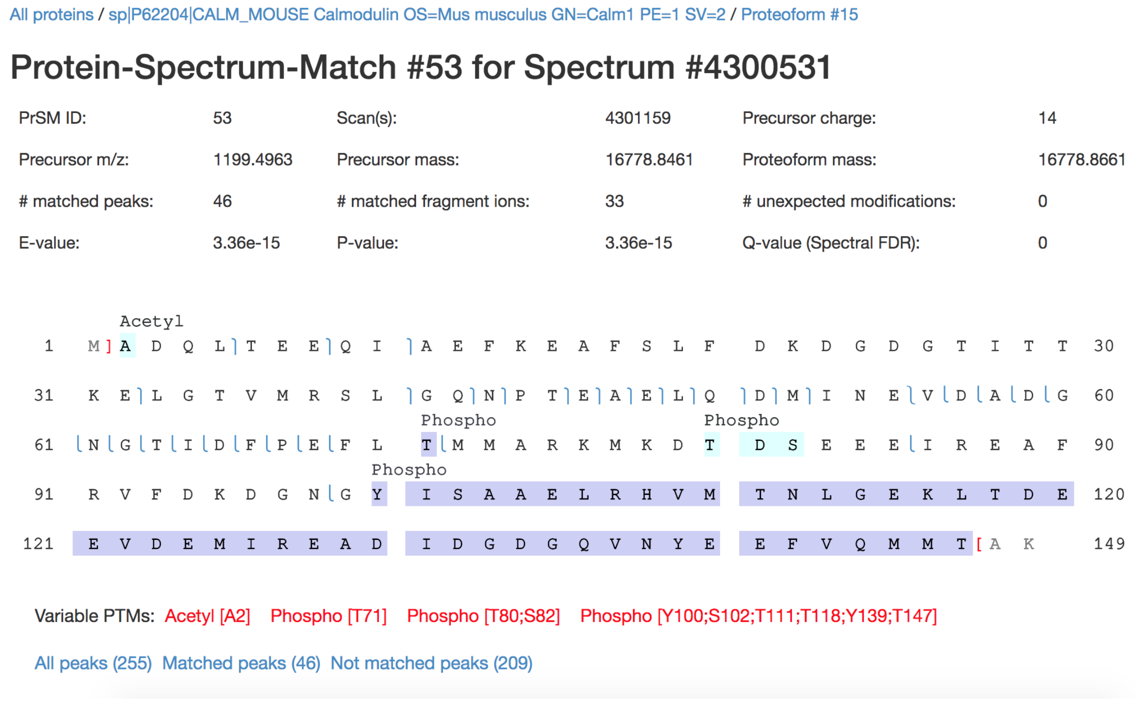
PrSM ID:	48	Scan(s):	4200516	Precursor charge:	18
Precursor m/z:	768.1483	Precursor mass:	13808.5382	Proteoform mass:	13808.8132
# matched peaks:	41	# matched fragment ions:	29	# unexpected modifications:	0
E-value:	2.50e-17	P-value:	2.50e-17	Q-value (Spectral FDR):	0

```
                                                         3 Phosphos
  1   M  P  D]P  A  K  S  A⌉P  A  ⌉P  K  K  G  S  K⌉K⌉A  V  T   K  V  Q  K  K  D  G  K  K  R   30

 31   K  R  S  R  K  E  S  Y  S  V   Y  V  Y  K⌊V⌊L  K  Q  V  H  ⌊P  D⌊T  G  I  S  S  K  A  M   60

 61  ⌊G  I⌊M⌊N⌊S⌊F⌊V⌊N⌊D⌊I   F  E  R  I  A  G  E  A  S  R   L  A  H  Y  N  K  R  S  T  I   90

 91   T  S  R  E  I  Q  T  A  V  R   L  L  L  P  G  E⌊L  A  K⌊H  ⌊A⌊V⌊S  E⌊G  T⌊K⌊A⌊V  T  120

121  ⌊K⌊Y⌊T  S  S  K                                                                         126
```

Variable PTMs: 3 Phosphos [T20;S33;S37;Y38;S39;Y41;Y43]

All peaks (230) Matched peaks (41) Not matched peaks (189)

Figure S24: A phosphorylated proteoform of the mouse histone H2B type 2-B with three phosphorylation sites (UniProt ID: Q64525) identified by TopMG.



All proteins / sp|P62204|CALM_MOUSE Calmodulin OS=Mus musculus GN=Calm1 PE=1 SV=2 / Proteoform #15

Protein-Spectrum-Match #53 for Spectrum #4300531

PrSM ID:	53	Scan(s):	4301159	Precursor charge:	14
Precursor m/z:	1199.4963	Precursor mass:	16778.8461	Proteoform mass:	16778.8661
# matched peaks:	46	# matched fragment ions:	33	# unexpected modifications:	0
E-value:	3.36e-15	P-value:	3.36e-15	Q-value (Spectral FDR):	0

```
        Acetyl
  1   M]A  D  Q  L⌉T  E  E⌉Q  I  ⌉A  E  F  K  E  A  F  S  L  F   D  K  D  G  D  G  T  I  T  T   30

 31   K  E⌉L  G  T  V  M  R  S  L  ⌉G  Q⌉N⌉P  T⌉E⌉A⌉E⌉L⌉Q  ⌉D⌉M⌉I  N  E⌊V⌊D⌊A⌊D⌊G   60
                                Phospho                          Phospho
 61  ⌊N⌊G⌊T⌊I⌊D⌊F⌊P⌊E⌊F  L   T⌊M  M  A  R  K  M  K  D  T   D  S  E  E  E⌊I  R  E  A  F   90
                             Phospho
 91   R  V  F  D  K  D  G  N⌊G  Y   I  S  A  A  E  L  R  H  V  M   T  N  L  G  E  K  L  T  D  E  120

121   E  V  D  E  M  I  R  E  A  D   I  D  G  D  G  Q  V  N  Y  E   E  F  V  Q  M  M  T[A  K     149
```

Variable PTMs: Acetyl [A2] Phospho [T71] Phospho [T80;S82] Phospho [Y100;S102;T111;T118;Y139;T147]

All peaks (255) Matched peaks (46) Not matched peaks (209)

Figure S25: A phosphorylated proteoform of the mouse calmodulin protein (UniProt ID: P62204) with three phosphorylation sites identified by TopMG.

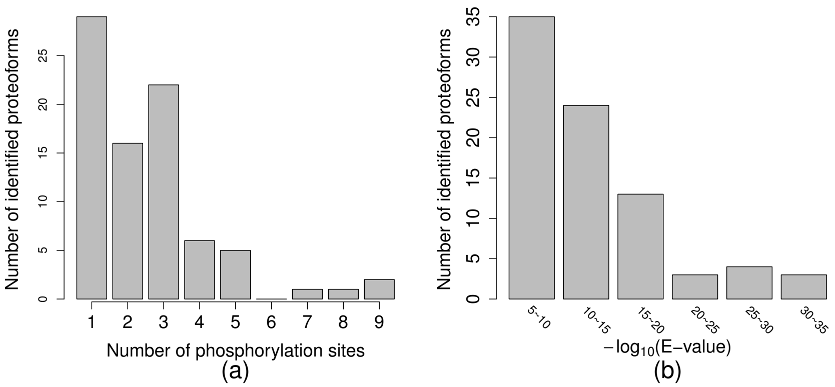
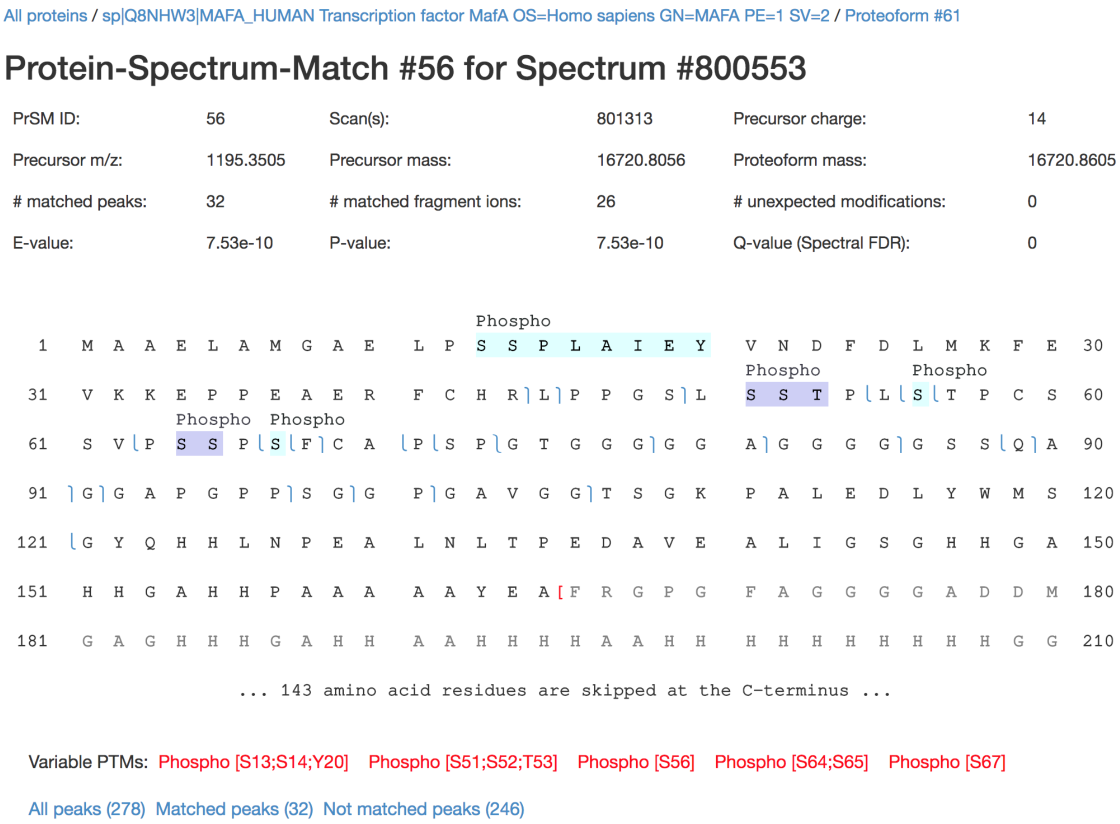


Figure S26: Histograms for the 82 phosphorylated human proteoforms identified from the xenograft data set by TopMG: (a) the number of phosphorylation sites, (b) E-values.



All proteins / sp|Q8NHW3|MAFA_HUMAN Transcription factor MafA OS=Homo sapiens GN=MAFA PE=1 SV=2 / Proteoform #61

Protein-Spectrum-Match #56 for Spectrum #800553

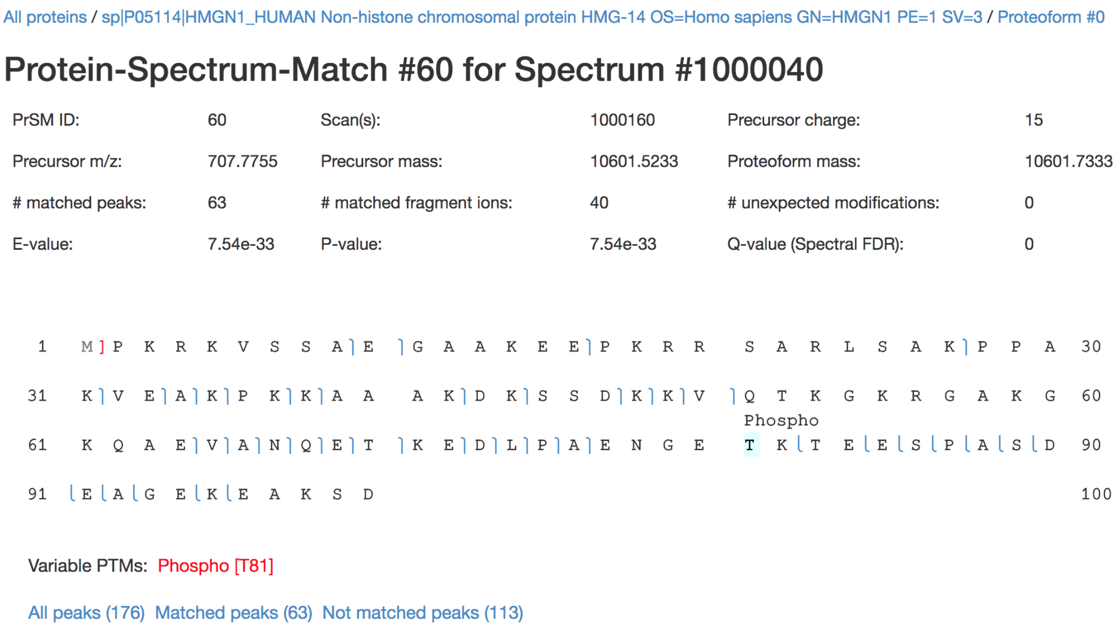
PrSM ID:	56	Scan(s):	801313	Precursor charge:	14
Precursor m/z:	1195.3505	Precursor mass:	16720.8056	Proteoform mass:	16720.8605
# matched peaks:	32	# matched fragment ions:	26	# unexpected modifications:	0
E-value:	7.53e-10	P-value:	7.53e-10	Q-value (Spectral FDR):	0

```
                                 Phospho
  1   M A A E L A M G A E   L P S S P L A I E Y   V N D F D L M K F E   30
                                                  Phospho     Phospho
 31   V K K E P P E A E R   F C H R⌉L⌉P P G S⌉L   S S T P⌊L⌊S⌊T P C S   60
              Phospho Phospho
 61   S V⌊P S S P⌊S⌊F⌉C A  ⌊P⌊S P⌉G T G G G⌉G G   A⌉G G G G⌉G S S⌊Q⌉A   90
 91  ⌉G⌉G A P G P P⌉S G⌉G   P⌉G A V G G⌉T S G K   P A L E D L Y W M S  120
121  ⌊G Y Q H H L N P E A   L N L T P E D A V E   A L I G S G H H G A  150
151   H H G A H H P A A A   A A Y E A[F R G P G   F A G G G G A D D M  180
181   G A G H H H G A H H   A A H H H H A A H H   H H H H H H H H G G  210
            ... 143 amino acid residues are skipped at the C-terminus ...
```

Variable PTMs: Phospho [S13;S14;Y20] Phospho [S51;S52;T53] Phospho [S56] Phospho [S64;S65] Phospho [S67]

All peaks (278) Matched peaks (32) Not matched peaks (246)

Figure S27: A phosphorylated proteoform of the human transcription factor MafA (UniProt ID: Q8NHW3) with five phosphorylation sites identified by TopMG.



All proteins / sp|P05114|HMGN1_HUMAN Non-histone chromosomal protein HMG-14 OS=Homo sapiens GN=HMGN1 PE=1 SV=3 / Proteoform #0

Protein-Spectrum-Match #60 for Spectrum #1000040

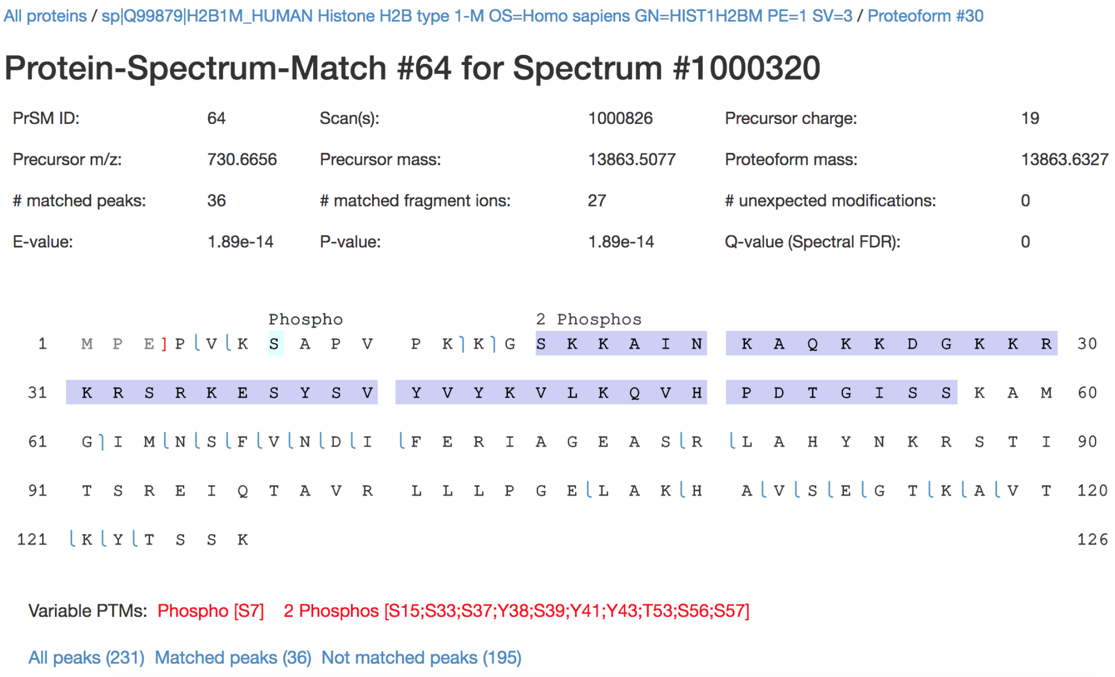
PrSM ID:	60	Scan(s):	1000160	Precursor charge:	15
Precursor m/z:	707.7755	Precursor mass:	10601.5233	Proteoform mass:	10601.7333
# matched peaks:	63	# matched fragment ions:	40	# unexpected modifications:	0
E-value:	7.54e-33	P-value:	7.54e-33	Q-value (Spectral FDR):	0

```
 1   M⌉P  K  R  K  V  S  S  A⌉E   ⌉G  A  A  K  E  E⌉P  K  R  R    S  A  R  L  S  A  K⌉P  P  A   30

31   K⌉V  E⌉A⌉K⌉P  K⌉K⌉A  A    A  K⌉D  K⌉S  S  D⌉K⌉K⌉V   ⌉Q  T  K  G  K  R  G  A  K  G   60
                                                         Phospho
61   K  Q  A  E⌉V⌉A⌉N⌉Q⌉E⌉T   ⌉K  E⌉D⌉L⌉P⌉A⌉E  N  G  E    T  K⌊T  E⌊E⌊S⌊P⌊A⌊S⌊D   90

91  ⌊E⌊A⌊G  E⌊K⌊E  A  K  S  D                                                                100
```

Variable PTMs: Phospho [T81]

All peaks (176) Matched peaks (63) Not matched peaks (113)

Figure S28: A phosphorylated proteoform of the human non-histone chromosomal protein HMG-14 (UniProt ID: P05114) with one phosphorylation site identified by TopMG.



All proteins / sp|Q99879|H2B1M_HUMAN Histone H2B type 1-M OS=Homo sapiens GN=HIST1H2BM PE=1 SV=3 / Proteoform #30

Protein-Spectrum-Match #64 for Spectrum #1000320

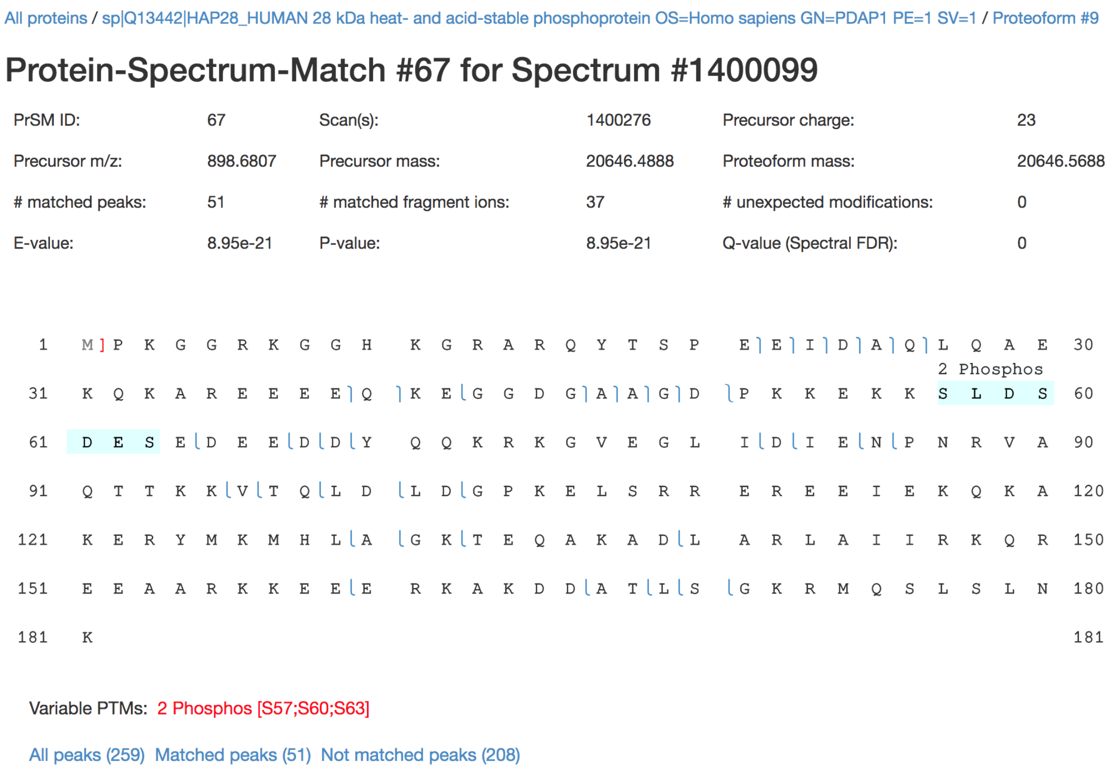
PrSM ID:	64	Scan(s):	1000826	Precursor charge:	19
Precursor m/z:	730.6656	Precursor mass:	13863.5077	Proteoform mass:	13863.6327
# matched peaks:	36	# matched fragment ions:	27	# unexpected modifications:	0
E-value:	1.89e-14	P-value:	1.89e-14	Q-value (Spectral FDR):	0

```
                      Phospho                 2 Phosphos
  1   M  P  E⌉P⌊V⌊K  S  A  P  V    P  K⌉K⌉G  S  K  K  A  I  N    K  A  Q  K  K  D  G  K  K  R   30

 31   K  R  S  R  K  E  S  Y  S  V    Y  V  Y  K  V  L  K  Q  V  H    P  D  T  G  I  S  S  K  A  M   60

 61   G⌉I  M⌊N⌊S⌊F⌊V⌊N⌊D⌊I   ⌊F  E  R  I  A  G  E  A  S⌊R   ⌊L  A  H  Y  N  K  R  S  T  I   90

 91   T  S  R  E  I  Q  T  A  V  R    L  L  L  P  G  E⌊L  A  K⌊H    A⌊V⌊S⌊E⌊G  T⌊K⌊A⌊V  T  120

121  ⌊K⌊Y⌊T  S  S  K                                                                          126
```

Variable PTMs: Phospho [S7] 2 Phosphos [S15;S33;S37;Y38;S39;Y41;Y43;T53;S56;S57]

All peaks (231) Matched peaks (36) Not matched peaks (195)

Figure S29: A phosphorylated proteoform of the human histone H2B type 1-M (UniProt ID: Q99879) with three phosphorylation sites identified by TopMG.



All proteins / sp|Q13442|HAP28_HUMAN 28 kDa heat- and acid-stable phosphoprotein OS=Homo sapiens GN=PDAP1 PE=1 SV=1 / Proteoform #9

Protein-Spectrum-Match #67 for Spectrum #1400099

PrSM ID:	67	Scan(s):	1400276	Precursor charge:	23
Precursor m/z:	898.6807	Precursor mass:	20646.4888	Proteoform mass:	20646.5688
# matched peaks:	51	# matched fragment ions:	37	# unexpected modifications:	0
E-value:	8.95e-21	P-value:	8.95e-21	Q-value (Spectral FDR):	0

```
  1   M⌉P  K  G  G  R  K  G  G  H    K  G  R  A  R  Q  Y  T  S  P    E⌉E⌉I⌉D⌉A⌉Q⌉L  Q  A  E   30
                                                                      2 Phosphos
 31   K  Q  K  A  R  E  E  E  E⌉Q   ⌉K  E⌊G  G  D  G⌉A⌉A⌉G⌉D   ⌊P  K  K  E  K  K  S  L  D  S   60
 61   D  E  S  E⌊D  E  E⌊D⌊D⌊Y    Q  Q  K  R  K  G  V  E  G  L    I⌊D⌊I  E⌊N⌊P  N  R  V  A   90
 91   Q  T  T  K  K⌊V⌊T  Q⌊L  D   ⌊L  D⌊G  P  K  E  L  S  R  R    E  R  E  E  I  E  K  Q  K  A  120
121   K  E  R  Y  M  K  M  H  L⌊A   ⌊G  K⌊T  E  Q  A  K  A  D⌊L    A  R  L  A  I  I  R  K  Q  R  150
151   E  E  A  A  R  K  K  E  E⌊E    R  K  A  K  D  D⌊A  T⌊L⌊S   ⌊G  K  R  M  Q  S  L  S  L  N  180
181   K                                                                                           181
```

Variable PTMs: 2 Phosphos [S57;S60;S63]

All peaks (259) Matched peaks (51) Not matched peaks (208)

Figure S30: A phosphorylated proteoform of the human 28 kDa heat- and acid-stable phosphoprotein (UniProt ID: Q13442) with two phosphorylation sites identified by TopMG.

Supplementary tables (Microsoft Excel files)

Table S4: A total of 874 PrSMs were identified from the EC data set with a 1% spectrum level FDR by TopPIC.

Table S8: A total of 3205 proteoforms were identified from the histone H3 data set with at least 10 matched fragment ions by TopMG.

Table S9: A total of 1087 proteoforms were identified from the histone H4 data set with at least 10 matched fragment ions by TopMG.

Table S12: A total of 122 mouse proteoforms were identified from the xenograft data set with a 5% proteoform-level FDR by TopPIC.

Table S13: A total of 45 mouse proteoforms were identified from the xenograft data set with a 5% proteoform-level FDR by TopMG.

Table S14: A total of 41 phosphorylated mouse proteoforms were identified from the xenograft data

set with a 5% proteoform-level FDR by TopMG.

Table S15: A total of 265 human proteoforms were identified from the xenograft data set with a 5% proteoform-level FDR by TopPIC.

Table S16: A total of 91 human proteoforms were identified from the xenograft data set with a 5% proteoform-level FDR by TopMG.

Table S17: A total of 82 phosphorylated human proteoforms were identified from the xenograft data set with a 5% proteoform-level FDR by TopMG.

Supplementary files (html files for annotated PrSMs)

File_S1_mouse_proteoforms_identified_by_TopPIC.zip

File_S2_mouse_proteoforms_identified_by_TopMG.zip

File_S3_human_proteoforms_identified_by_TopPIC.zip

File_S4_human_proteoforms_identified_by_TopMG.zip

References

[1] Leonid Zamdborg, Richard D LeDuc, Kevin J Glowacz, Yong-Bin Kim, Vinayak Viswanathan, Ian T Spaulding, Bryan P Early, Eric J Bluhm, Shannee Babai, and Neil L Kelleher. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research*, 35(Web Server issue):W701–W706, 2007.

[2] Ari M Frank, James J Pesavento, Craig A Mizzen, Neil L Kelleher, and Pavel A Pevzner. Interpreting top-down mass spectra using spectral alignment. *Analytical Chemistry*, 80:2499–2505, 2008.

[3] Yihsuan S Tsai, Alexander Scherl, Jason L Shaw, C. Logan MacKay, Scott A Shaffer, Patrick R R Langridge-Smith, and David R Goodlett. Precursor ion independent algorithm for top-down shotgun proteomics. *Journal of the American Society for Mass Spectrometry*, 20:2154–2166, 2009.

[4] N. Murat Karabacak, Long Li, Ashutosh Tiwari, Lawrence J Hayward, Pengyu Hong, Michael L Easterling, and Jeffrey N Agar. Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Molecular & Cellular Proteomics*, 8:846–856, 2009.

[5] Weiwei Tong, Roger Théberge, Giuseppe Infusini, David H Perlman, Catherine E Costello, and Mark E McComb. BUPID-top-down: database search and assignment of top-down MS/MS data. In *Proceedings of the 57th American Society Conference on Mass Spectrometry and Allied Topics, Philadelphia, PA*, volume 31, 2009.

[6] Xiaowen Liu, Yakov Sirotkin, Yufeng Shen, Gordon Anderson, Yihsuan S Tsai, Ying S Ting, David R Goodlett, Richard D Smith, Vineet Bafna, and Pavel A Pevzner. Protein identification using top-down spectra. *Molecular & Cellular Proteomics*, 11:M111.008524, 2012.

[7] M. Bern, Y. J. Kil, and C. Becker. Byonic: advanced peptide and protein identification software. *Current Protocols in Bioinformatics*, Chapter 13:Unit 13.20, 2012.

[8] Xiaowen Liu, Shawna Hengel, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, and Pavel A Pevzner. Identification of ultramodified proteins using top-down tandem mass spectra. *Journal of Proteome Research*, 12:5830–5838, 2013.

[9] Li Li and Zhixin Tian. Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. *Rapid Communications in Mass Spectrometry*, 27:1267–1277, 2013.

[10] Kaijie Xiao, Fan Yu, Houqin Fang, Bingbing Xue, Yan Liu, and Zhixin Tian. Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. *Scientific Reports*, 5, 2015.

[11] Ryan T Fellers, Joseph B Greer, Bryan P Early, Xiang Yu, Richard D LeDuc, Neil L Kelleher, and Paul M Thomas. ProSight Lite: graphical software to analyze top-down mass spectrometry data. *Proteomics*, 15:1235–1238, 2015.

[12] Michael R Shortreed, Brian L Frey, Mark Scalf, Rachel A Knoener, Anthony J Cesnik, and Lloyd M Smith. Elucidating proteoform families from proteoform intact-mass and lysine-count measurements. *Journal of Proteome Research*, 15:1213–1221, 2016.

[13] Qiang Kou, Likun Xun, and Xiaowen Liu. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, 32:3495–3497, 2016.

[14] R. X. Sun, L. Luo, L. Wu, R. M. Wang, W. F. Zeng, H. Chi, C. Liu, and S. M. He. pTop 1.0: A high-accuracy and high-efficiency search engine for intact protein identification. *Analytical Chemistry*, 88:3082–90, 2016.

[15] Qiang Kou, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, Yunlong Liu, and Xiaowen Liu. A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. *Bioinformatics*, 33:1309–1316, 2016.

[16] Huseyin Guner, Patrick L Close, Wenxuan Cai, Han Zhang, Ying Peng, Zachery R Gregorich, and Ying Ge. MASH Suite: a user-friendly and versatile software interface for high-resolution mass

spectrometry data interpretation and visualization. *Journal of the American Society for Mass Spectrometry*, 25:464–470, 2014.

[17] Wenxuan Cai, Huseyin Guner, Zachery R Gregorich, Albert J Chen, Serife Ayaz-Guner, Ying Peng, Santosh G Valeja, Xiaowen Liu, and Ying Ge. MASH Suite Pro: A comprehensive software tool for top-down proteomics. *Molecular & Cellular Proteomics*, 15:703–714, 2016.

[18] Jungkap Park, Paul D Piehowski, Christopher Wilkins, Mowei Zhou, Joshua Mendoza, Grant M Fujimoto, Bryson C Gibbons, Jared B Shaw, Yufeng Shen, Anil K Shukla, Ronald J Moore, Tao Liu, Vladislav A Petyuk, Nikola Tolić, Ljiljana Paša-Tolić, Richard D Smith, Samuel H Payne, and Sangtae Kim. Informed-Proteomics: open-source software package for top-down proteomics. *Nature Methods*, 14:909–914, 2017.

[19] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11:320–332, 2000.

[20] Vlad Zabrouskov, Michael W Senko, Yi Du, Richard D Leduc, and Neil L Kelleher. New and automated MSn approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry*, 16:2027–2038, 2005.

[21] Anoop M Mayampurath, Navdeep Jaitly, Samuel O Purvine, Matthew E Monroe, Kenneth J Auberry, Joshua N Adkins, and Richard D Smith. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, 24:1021–1023, 2008.

[22] Paulo C Carvalho, Tao Xu, Xuemei Han, Daniel Cociorva, Valmir C Barbosa, and John R Yates III. YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics*, 25:2734–2736, 2009.

[23] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, and R. D. Smith. Decon2ls: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, 10:87, 2009.

[24] Xiaowen Liu, Yuval Inbar, Pieter C Dorrestein, Colin Wynne, Nathan Edwards, Puneet Souda, Julian P Whitelegge, Vineet Bafna, and Pavel A Pevzner. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Molecular & Cellular Proteomics*, 9:2772–2782, 2010.

[25] Qiang Kou, Si Wu, and Xiaowen Liu. A new scoring function for top-down spectral deconvolution. *BMC Genomics*, 15:1140, 2014.

[26] Z. F. Yuan, C. Liu, H. P. Wang, R. X. Sun, Y. Fu, J. F. Zhang, L. H. Wang, H. Chi, Y. Li, L. Y. Xiu, W. P. Wang, and S. M. He. pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics*, 12:226–35, 2012.

[27] M. T. Marty, A. J. Baldwin, E. G. Marklund, G. K. Hochberg, J. L. Benesch, and C. V. Robinson. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical Chemistry*, 87:4370–6, 2015.

[28] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77:4626–39, 2005.

[29] MS-Align+Tag. `http://bioinf.spbau.ru/proteomics/ms-align-plus-tag`, 2012.