SUPPORTING INFORMATION

Constructing Human Proteoform Families Using Intact-Mass and Top-Down Proteomics with a Multi-Protease Global PTM Discovery Database

Yunxiang Dai,[1,2,#] Katherine E. Buxton,[1,#] Leah V. Schaffer,[1] Rachel M. Miller,[1] Robert J. Millikin,[1] Mark Scalf,[1] Brian L. Frey,[1] Michael R. Shortreed,[1] Lloyd M. Smith[*,1]

[1] Department of Chemistry, University of Wisconsin, Madison, Wisconsin, United States
[2] Biophysics Graduate Program, University of Wisconsin, Madison, Wisconsin, United States
[#] These authors contributed equally to this work.
[*]Corresponding author: E-mail: smith@chem.wisc.edu. Phone: (608) 263-2594. Fax: (608) 265-6780.

Table of Contents:

SUPPLEMENTARY EXPERIMENTAL METHODS

Materials

- Jurkat cells (TIB-152) were purchased from the American Type Culture Collection (ATCC) (Manassas, VA).
- SILAC RPMI-1640 medium (A2494401), fetal bovine serum (26400036), antibiotic-antimycotic solution (15240062), HEPES buffer (15630080), sodium pyruvate solution (11360070), GlutaMAX (35050061), 100X HALT protease/phosphatase inhibitor cocktail (78441), and methanol (A452) were purchased from Thermo Fisher Scientific (Waltham, MA).
- L-arginine (A5006), DL-dithiothreitol (DTT) (D5545), sodium butyrate (B5887), iodoacetamide (I1149), acetone (270725), unlabeled L-lysine (62840), 10% sodium dodecyl sulfate (SDS) (71736), and chloroform (319988) were purchased from Sigma-Aldrich Corp. (St. Louis, MO).
- L-lysine:2HCl, $^{13}C_6{}^{15}N_2$ (CNLM-291-H) and L-lysine:2HCl, 3,3,4,4,5,5,6,6 D_8 (DLM-2641) were purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA).
- 4 M Tris-HCl pH=7.5 (T5575), 10X phosphate buffered saline (PBS) (P0195), and 500 mM EDTA pH=8.0 (E0306) were purchased from Teknova (Hollister, CA).
- Gelfree 8100 12% Tris-acetate cartridge (42402), Tris-acetate 5X sample buffer (42302), and HEPES running buffer (42202) were purchased from Expedeon (San Diego, CA).
- Acetonitrile (ACN) (AH015-4) was purchased from Honeywell (Morris Plains, NJ).
- Formic acid (11670) was purchased from EMD Millipore (Burlington, MA).

Cell Lysis for NeuCode Intact-Mass Proteomics

For each replicate NeuCode intact-mass experiment, approximately 10^7 "light" and 10^7 "heavy" labeled Jurkat cells were thawed and lysed separately in 1 mL buffer containing 4% (w/v) SDS, 100 mM Tris-HCl pH = 7.5, 10 mM DTT, 10 mM sodium butyrate, 20 mM EDTA, and 1X HALT protease/phosphatase inhibitors. The cells were incubated at room temperature for 10 min with frequent vortexing, then sonicated in a water bath sonicator (FS20, Fisher Scientific) for 5 min with 20 s on/off intervals. The lysate was incubated for an additional 30 min at room temperature, then proteins were alkylated with 20 mM iodoacetamide for 30 min. Residual iodoacetamide was quenched via a 15 min incubation with a final concentration of 20 mM DTT. Proteins were precipitated with acetone at -20°C and resuspended in 200 µL of 1% SDS. Proteins from the two NeuCode-labeled samples were then mixed in a 2:1 "light":"heavy" ratio by volume.

Gelfree Fractionation

Two ~65 µL aliquots of this mixed protein sample were added to new 1.7 mL tubes. Sample buffer (30 µL), 1 M DTT (8 µL), and water were added to bring the total volume in each tube up to 150 µL. The tubes were incubated at 50 ˚C for 10 min, then cooled to room temperature. The contents of each tube were fractionated in separate 12% Tris-acetate Gelfree cartridge channels using the manufacturer's recommended procedure. To prepare the channels, storage buffer was removed and replaced with running buffer. Each of the 150 µL samples was loaded, and a standard running method was used to separate the samples into fractions based on molecular weight. Between each step in the method, fractions in the collection chambers of the two channels were combined into a new 2 mL low-retention tube. The collection chambers were rinsed and replenished with new running buffer. These steps were repeated 11 times for each fraction collection. Throughout the run, the running buffer was changed twice according to the standard procedure. The collected fractions were stored at −20°C or 4°C for subsequent sample preparation. Prior to mass spectrometric analysis, SDS was removed from each of the fractions via methanol-chloroform precipitation[1] and proteins were reconstituted with 16 µL of 5% ACN and 0.2% formic acid in water. Intact protein solutions were gently vortexed and centrifuged on a bench-top

centrifuge for 1 min. Solutions were carefully transferred into HPLC sample vials, leaving behind undissolved substances. Three biological replicates of this experiment were performed.

LC/MS Methods for NeuCode Intact-Mass Proteomics

All fractions were analyzed by HPLC-ESI-MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). HPLC separation employed a 100 × 365 μm fused silica capillary microcolumn packed with 20 cm of 5 μm diameter, 1000 Å pore size PLRP-S resin (Agilent) with an emitter tip pulled to approximately 1 μm using a laser puller (Sutter Instruments). Proteins were loaded on-column at a flow rate of 500 nL/min for 30 min, then eluted at 500 nL/min over 67 min with a gradient of 5% to 85% ACN in 0.2% formic acid. Full-mass profile scans were performed between 500 and 1,600 *m/z* at a resolution of 240,000. Seven microscans were averaged, using an AGC target of 3×10^6 with a maximum injection time of 200 ms. Source-induced dissociation was set to 15.0 eV. Two technical replicate injections of each fraction were performed, yielding a total of 66 raw data files (3 biological replicates × 11 fractions × 2 injections).

Bottom-Up Proteomics

Bottom-up proteomics data were collected previously for five aliquots of Jurkat cell lysate, each of which was digested with a different protease (chymotrypsin, GluC, ArgC, AspN, and LysC).[2] Methods for cell culture, lysis, FASP (including different digestion conditions for different proteases), peptide fractionation via high-pH reversed-phase liquid chromatography, and LC/MS are described in detail elsewhere.[2-3] As part of a separate work, this process was repeated to collect MS data for tryptic peptides from unlabeled Jurkat cell lysate.[4] The only alterations to the aforementioned procedure were as follows: proteins were digested in 50 mM ammonium bicarbonate buffer (pH = 7.8) using a 1:50 trypsin:protein ratio, 10 fractions of peptides were collected (instead of 11), and peptides were reconstituted with 5% ACN/1% formic acid in water prior to LC/MS.

Gelfree Fractionation and LC/MS Methods for Top-Down Proteomics

Proteins from label-free Jurkat cell lysate were fractionated by Gelfree as described for the NeuCode-labeled samples, except ~110 μL of resuspended protein were fractionated in a single Gelfree channel. Top-down analysis of each of the 11 Gelfree fractions was performed via HPLC−ESI−MS/MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). The LC method was the same as the intact-mass experiments. MS1 scans were performed between 500 and 1,600 *m/z* at a resolution of 240,000. Seven microscans were averaged, using an AGC target of 1×10^6 with a maximum injection time of 100 ms. The top three most intense peaks in the MS1 with $z > 2$ were selected for HCD fragmentation with a normalized collision energy setting of 25. The MS2 resolution was 120,000, the isolation window was 4 *m/z* units, and three microscans were averaged. Dynamic exclusion was enabled with a duration of 30 s. Source-induced dissociation was set to 15.0 eV. One biological replicate and two technical replicates were performed for this analysis, generating 22 raw data files.

Intact-Mass Data Deconvolution

Intact-mass raw files were deconvoluted into monoisotopic components using Protein Deconvolution 4.0 software (Thermo Fisher Scientific) (minimum S/N = 2, minimum number of detected charge states = 2, fit factor = 70%, remainder threshold = 10%, target average spectrum width = 0.18 min, target average spectrum offset = 34%). Different charge state ranges were selected for deconvoluting different fractions: +5 to +30 for fractions 1–9 and +5 to +50 for fractions 10–11. Each raw file was split into three to nine retention time (RT) ranges for deconvolution so that the output tables did not exceed allowed spreadsheet size. In the subsequent mass calibration process, calibrated deconvolution files for fractions 1 and 2 were not obtained due to insufficient data points. Therefore, fractions 1 and 2 were not analyzed further.

Proteoform Family Visualization with Description of Alterations
The proteoform family figures shown in the main manuscript (Figures 4 and 6) were modified slightly from the default Cytoscape output to eliminate false connections and improve clarity. In the Figure 4 bottom-right family (L28 family built with multi-protease G-PTM-D database), an edge (32 Da) between the 15,760.9 Da and 15,792.8 Da proteoforms was removed, as the new annotation of the 15,792.8 Da proteoform with the multi-protease G-PTM-D database is no longer a doubly-oxidized version of the 15,760.9 Da proteoform. In Figure 6 family A, a 10,720.8 Da proteoform was removed, as it was connected to the 10,752.8 Da proteoform with a 32 Da mass difference, but the latter did not contain two oxidations. In families A, C, and D, the edges between gene names and top-down proteoforms were removed for clarity. In both Figures 4 and 6, the slope of the linear function of node size vs. proteoform intensity has been adjusted for example families to enhance contrast.

*Previous publications by the Smith Group also contain helpful descriptions of experimental methods for proteoform identification.[5-9]

RAW MS DATA FILES AND G-PTM-D DATABASES

All raw data files and G-PTM-D databases are available on the MassIVE platform (MSV000083768, ftp://massive.ucsd.edu/MSV000083768, and MSV000083304, ftp://massive.ucsd.edu/MSV000083304). There are 66 MS files (in .raw format) from NeuCode-labeled intact-mass proteomics (11 Gelfree fractions of 3 biological replicates with 2 technical replicate injections each). There are 22 MS files from label-free top-down proteomics (11 fractions of one biological replicate with 2 technical replicate injections each). There are also 65 MS files from multi-protease bottom-up proteomics (10 fractions for trypsin and 11 fractions for each of the other five proteases). The pruned trypsin-only G-PTM-D database and the pruned multi-protease G-PTM-D database built from these bottom-up data are also included. The intact-mass data, top-down data, and G-PTM-D databases can be found in data set MSV000083768. The bottom-up data can be found in data set MSV000083304.

SUPPLEMENTARY RESULTS

NeuCode Intact-Mass-Only Analysis Using Unpruned Multi-Protease G-PTM-D Database
We performed a Proteoform Suite analysis with calibrated NeuCode intact-mass data but using the full multi-protease G-PTM-D database instead of the pruned database. The full G-PTM-D database contained a large number of protein sequences (~61,000) and PTMs that were not detected in our bottom-up data, which expanded the search space significantly. The full G-PTM-D database was used to construct a theoretical proteoform catalog that contained 426,170 entries, which is 3.5 times larger than pruned database-derived catalog. Using the same data filtering parameters and selecting the same ET and EE peaks, 561 unique proteoforms were identified in this analysis at 30% FDR (selecting peaks with the same low FDRs as those in the pruned database-assisted analysis was not possible). This high FDR is a result of the large search space of the theoretical proteoform catalog, which contributed to high FDRs in ET pairs (the highest being 134%). This analysis illustrates how using a pruned G-PTM-D database can help to prevent high FDRs during proteoform identification.

Integrated Intact-Mass/Top-Down Analysis Using Uncalibrated Data
We performed a Proteoform Suite analysis using uncalibrated intact-mass and top-down data with a catalog generated from the pruned multi-protease G-PTM-D database. Raw mass components from these uncalibrated data were filtered with the same parameters reported in the main manuscript. ET

comparisons revealed that only 103 pairs grouped to the 0.0185 Da peak (10% FDR), the only peak indicating exact matches. This is only 12% of the 863 exact-matching pairs in the analysis with calibrated data, leading to many fewer identifications in this analysis. Therefore, mass calibration is crucial to effectively identify proteoforms.

Top-Down-Only Analysis Utilizing Top-Down MS1 Spectra as "Label-Free Intact-Mass" Data
We attempted to maximize the utility of the proteomics data collected, so we delved deeper into our label-free top-down data to look for unidentified proteoforms observed in the MS1 spectra. This data analysis strategy was previously employed in yeast and murine mitochondrial proteoform analyses.[8-9]

Precursor ion spectra (MS1) of the top-down data files were extracted. These new top-down MS1-only files were referred to as "label-free intact-mass data files". They were deconvoluted like NeuCode intact-mass data in Thermo Protein Deconvolution 4.0 to provide label-free raw mass components. The resultant mass components were calibrated as described in the main manuscript. Processed intact-mass data were imported into Proteoform Suite together with calibrated top-down hits. A theoretical proteoform catalog was constructed using the same multi-protease G-PTM-D database and the "patch" databases as described in the main manuscript, except only one PTM was allowed instead of four to prevent high FDRs in the subsequent ET comparison stage. NeuCode pair determination and lysine count calculation were skipped, so ET and EE comparisons proceeded with aggregated proteoform masses. ET and EE pairs were grouped with smaller intervals (0.005 and 0.01 Da, respectively), so that mass difference peaks with relatively low FDRs would be revealed and accepted for family construction.

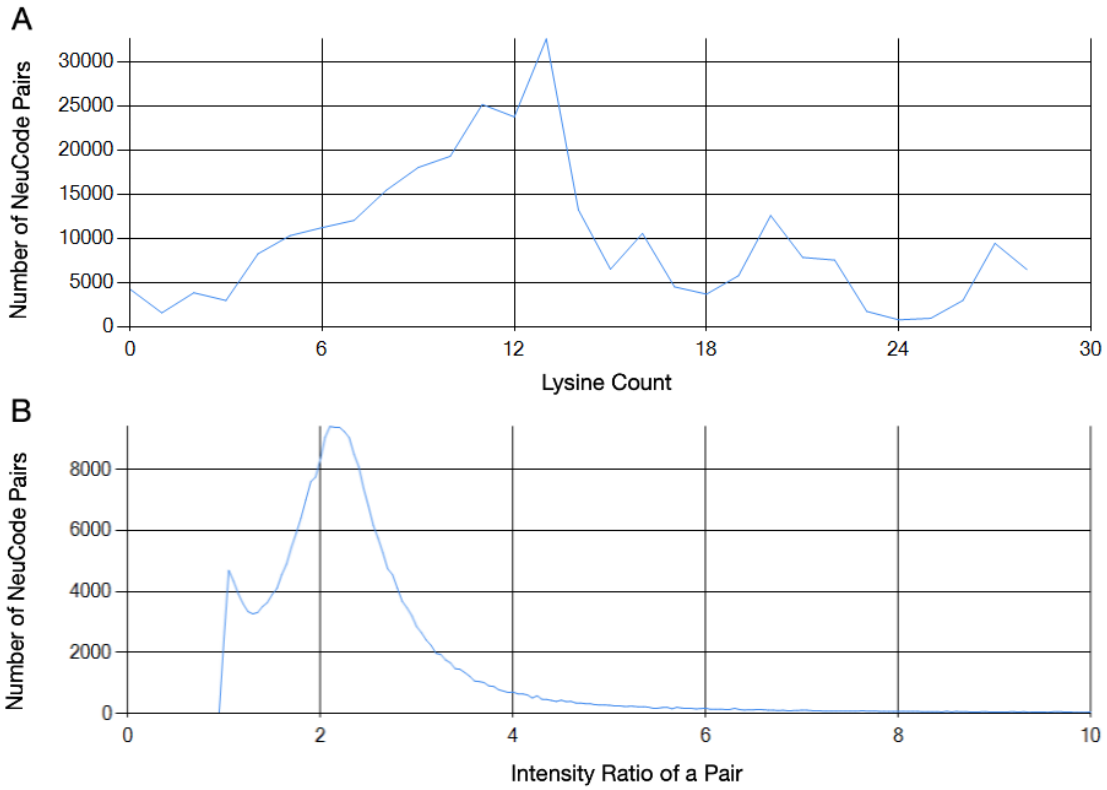
The label-free intact-mass data further increased the number of unique proteoform identifications beyond those revealed by the integrated NeuCode intact-mass/label-free top-down analysis discussed in the main manuscript. In this label-free intact-mass/top-down analysis, Proteoform Suite identified 880 proteoforms representing 444 genes. The overall FDR for proteoform identification was 5.4%, which is low considering the high noise level of ET and EE comparisons in label-free proteoform analysis (where we do not have the knowledge of lysine count to help limit the number of comparisons). We identified 169 unique intact-mass experimental proteoforms not found in the top-down hits. Among those, 120 proteoforms were new identifications not found in the NeuCode intact-mass proteoform identifications. These proteoforms represented 28 new genes not found in the previous analysis. These new identifications contributed to a 10% increase in total unique proteoform identifications, and a 5.8% increase in total genes represented, yielding a final result of 1,327 (1,207 + 120) proteoforms representing 512 (484 + 28) genes. Over 45% of the proteoform identifications came from intact-mass measurements (NeuCode-labeled and/or label-free).

Protein Entries in the Patch Database
The patch database contained 101 protein accession numbers. Fifty-six of these accessions had peptide-level evidence in the multi-protease bottom-up data, but the peptide of interest was observed above a 1% FDR, and therefore the accession was not included in the pruned multi-protease G-PTM-D database. Furthermore, 29 of the 101 accessions in the patch database were included as isoforms. The original UniProt XML database used to search the bottom-up data did not include isoform sequences, which explains why these 29 accessions were not included in the pruned multi-protease G-PTM-D database. That being said, peptide-level evidence for the canonical form of 26 of these proteins was observed in the bottom-up data (FDR > 2%). We posit that many of the remaining 16 accessions in the patch database were eliminated from the pruned multi-protease G-PTM-D database as a result of the protein inference process.

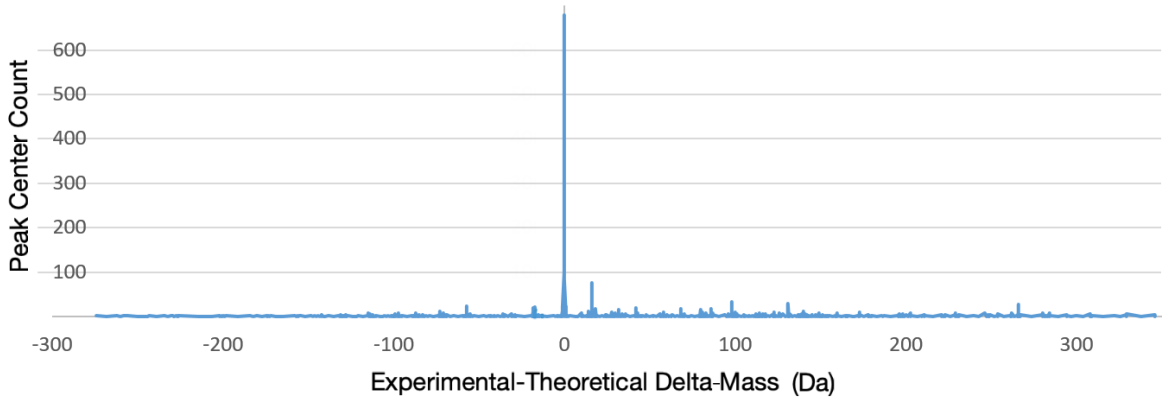
Histone Proteoforms and PTMs

Histone proteoforms are known to play important roles in gene expression. In this study, we identified 289 histone proteoforms among the 1,207 reported. These histone proteoforms contained PTMs including methylation, acetylation, phosphorylation, malonylation, succinylation, oxidation, deamidation, and carbamylation. Many of these PTM types play key roles in histone-mediated modulation of gene expression. There were 6 histone proteoform families assembled, including 5 ambiguously identified families and 1 unambiguously identified family (SI Table S-10). Ambiguous histone families account for half of the ambiguous families reported (5 out of 10), and this ambiguity is largely the result of the similar molecular weights of numerous types of histones H2A, H2B, and H3 (e.g., H2B type 1-A, B, C, etc.). It is important to note that, although most histone proteoforms were in the ambiguous families, many of them were still unambiguously identified through direct connections to known theoretical histone proteoforms. The one unambiguous histone family is the Histone H3.1t proteoform family, containing 3 top-down and 3 intact-mass experimental proteoforms (SI Figure S-6)

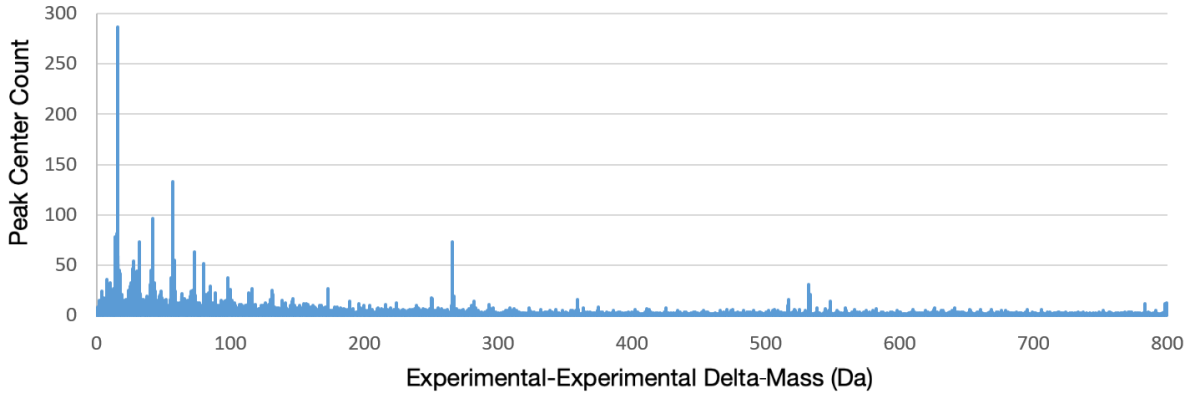
SUPPLEMENTARY FIGURES



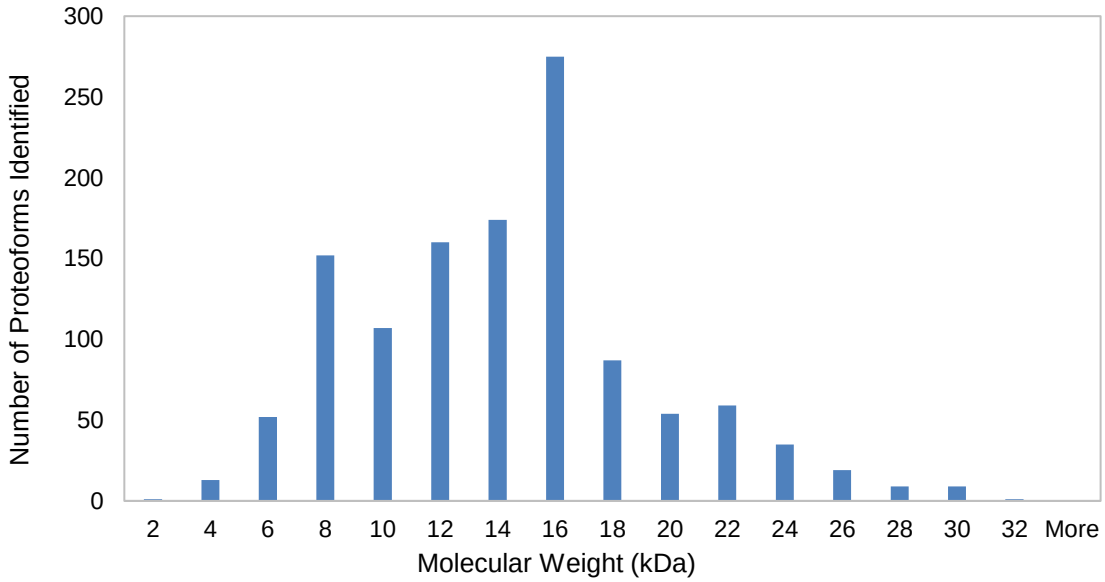
SI Figure S-1. Lysine count (A) and intensity ratio (B) distributions of the 283,634 NeuCode pairs that Proteoform Suite identified in this study. The peak intensity ratio was 2.15:1, which was close to the mixing ratio of "light" and "heavy" protein samples at 2:1. NeuCode pairs whose intensity ratios were between 1.8:1 to 2.5:1 were accepted for aggregation into experimental proteoforms.



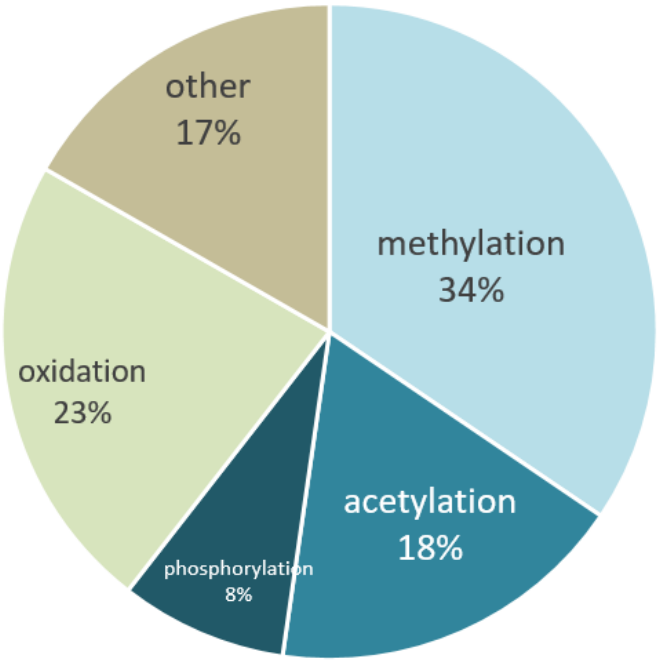
SI Figure S-2. Mass difference histogram from experimental-theoretical proteoform comparisons. Analysis was performed using intact-mass and top-down data with a catalog of theoretical proteoforms derived from the pruned multi-protease G-PTM-D database.



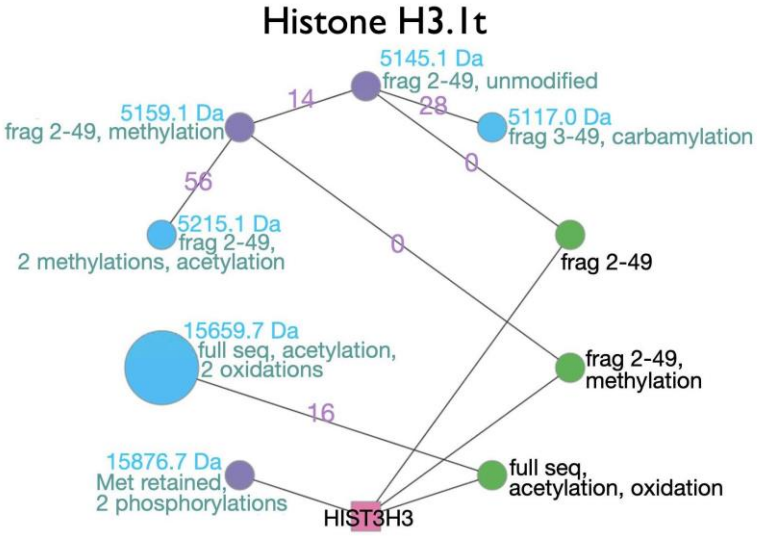
SI Figure S-3. Mass difference histogram from experimental-experimental proteoform comparisons. Analysis was performed using intact-mass and top-down data.



SI Figure S-4. Molecular weight distribution of the 1,207 identified proteoforms. This result is from the integrated intact-mass/top-down analysis.



SI Figure S-5. PTM types found in the 1,207 identified proteoforms. This result is from the integrated intact-mass/top-down analysis.



SI Figure S-6. Histone H3.1t proteoform family. Proteoforms in this family contain PTMs such as methylation, acetylation, and phosphorylation that are known to contribute to histone function.

REFERENCES FOR SUPPORTING INFORMATION

(1) Wessel, D.; Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **1984,** *138*, 141-143.
(2) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014,** *13*, 228-240.
(3) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-seq. *Mol. Cell. Proteomics* **2013,** *12*, 2341-2353.
(4) Miller, R. M.; Millikin, R. J.; Hoffman, C. V.; Solntsev, S. K.; Sheynkman, G. M.; Shortreed, M. R.; Smith, L. M. Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. *J. Proteome Res.* **2019**, *18*, 3429-3438.
(5) Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *J. Proteome Res.* **2016,** *15*, 1213-1221.
(6) Dai, Y.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Cesnik, A. J.; Solntsev, S.; Schaffer, L. V.; Smith, L. M. Elucidating Escherichia coli Proteoform Families Using Intact-Mass Proteomics and a Global PTM Discovery Database. *J. Proteome Res.* **2017,** *16*, 4156-4165.
(7) Cesnik, A. J.; Shortreed, M. R.; Schaffer, L. V.; Knoener, R. A.; Frey, B. L.; Scalf, M.; Solntsev, S. K.; Dai, Y.; Gasch, A. P.; Smith, L. M. Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. *J. Proteome Res.* **2018,** *17*, 568-578.
(8) Schaffer, L. V.; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; Scalf, M.; Smith, L. M. Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. *Anal. Chem.* **2018,** *90*, 1325-1333.
(9) Schaffer, L. V.; Rensvold, J. W.; Shortreed, M. R.; Cesnik, A. J.; Jochem, A.; Scalf, M.; Frey, B. L.; Pagliarini, D. J.; Smith, L. M. Identification and Quantification of Murine Mitochondrial Proteoforms Using an Integrated Top-Down and Intact-Mass Strategy. *J. Proteome Res.* **2018,** *17*, 3526-3536.