Title

Construction of Human Proteoform Families from 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Top-Down Proteomic Data

Journal

Journal of Proteome Research, 20(1)

ISSN

1535-3893

Authors

Schaffer, Leah V
Anderson, Lissa C
Butcher, David S
et al.

Publication Date

2021

DOI

10.1021/acs.jproteome.0c00403

Peer reviewed

HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

J Proteome Res. 2021 January 01; 20(1): 317–325. doi:10.1021/acs.jproteome.0c00403.

Construction of Human Proteoform Families from 21 Tesla FT-ICR Mass Spectrometry Top-Down Proteomic Data

Leah V. Schaffer[1], **Lissa C. Anderson**[2], **David S. Butcher**[2], **Michael R. Shortreed**[1], **Rachel M. Miller**[1], **Caitlin Pavelec**[1], **Lloyd M. Smith**[1,*]

[1]Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[2]Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Tallahassee, Florida 32310, United States

Abstract

Identification of proteoforms, the different forms of a protein, is important to understand biological processes. A proteoform family is the set of different proteoforms from the same gene. We previously developed the software program Proteoform Suite, which constructs proteoform families and identifies proteoforms by intact-mass analysis. Here, we have applied this approach to top-down proteomic data acquired at the National High Magnetic Field Laboratory 21 tesla FT-ICR mass spectrometer (data available on the MassIVE platform with identifier MSV000085978). We explored the ability to construct proteoform families and identify proteoforms from the high mass accuracy data that this instrument provides for a complex cell lysate sample from the MCF-7 human breast cancer cell line. 2830 experimental proteoforms were observed, of which 932 were identified, 44 were ambiguous, and 1854 were unidentified. Of the 932 unique identified proteoforms, 766 were identified by top-down MS2 analysis at 1% FDR using TDPortal and 166 were additional intact-mass identifications (~4.7% calculated global FDR) made using Proteoform Suite. We recently published a proteoform level schema to represent ambiguity in proteoform identifications. We implemented this proteoform level classification in Proteoform Suite for intact-mass identifications, which enables users to determine the ambiguity levels and sources of ambiguity for each intact-mass proteoform identification.

Graphical Abstract

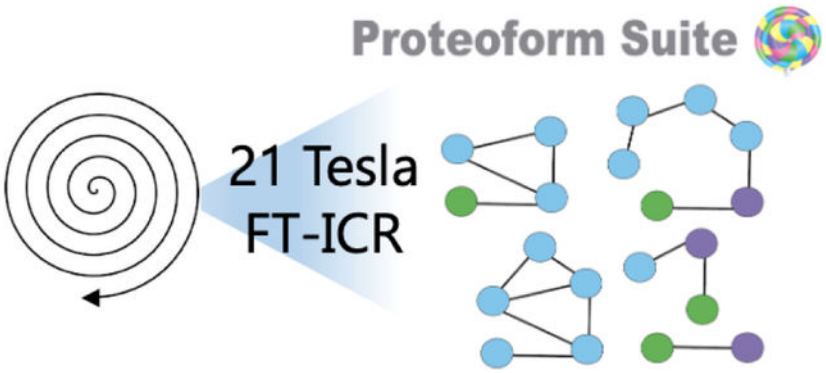
[*]**Corresponding Author**: smith@chem.wisc.edu. Phone: 608-263-2594. Fax: 608-265-6780.

Author Contributions

L.V.S. and M.R.S. developed Proteoform Suite. L.V.S. drafted the manuscript. L.C.A. and D.S.B. performed the mass spectrometry experiments. R.M.M. prepared the samples. L.V.S. and C.P. performed data analysis. L.M.S. provided oversight of the work. All authors edited the manuscript.

Supporting Information. Supporting Text: Retention Time Calibration, Bottom-Up MetaMorpheus Search, QE-HF Data Acquisition; Figure S-1: Top-down precursor mass error; Figure S-2: Top-down precursor retention time differences; Figure S-3: Experiment-theroetical and experiment-experiment delta mass histograms; Table S-1: Top-down proteoform identifications; Table S-2: Intact-mass proteoform identifications; Table S-3: Ambiguous intact-mass proteoforms; Table S-4: Unidentified intact-mass proteoforms; Table S-5: Top-Down Hit Results from QE-HF Orbitrap analysis of MCF-7; Table S-6: Intact-mass analysis of top-down precursor masses.

The authors declare no competing financial interests.



Keywords

Proteoform; proteoform family; 21 tesla; fourier-transform ion cyclotron resonance; top-down proteomics

INTRODUCTION

Protein diversity plays a central role in the function of biological systems.[1, 2] Proteoforms are the different forms of proteins that result from processes such as genetic variation, alternative splicing, and post-translational modification.[3] A proteoform family is the set of proteoforms derived from the same gene.[4] Proteoforms are identified by top-down mass spectrometry (MS), where intact proteins are analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) and the intact and fragment ion masses are used to identify and characterize the proteoform.[5–7] However, not all proteoforms observed in MS1 spectra are subsequently identified by MS2 fragmentation.[8, 9] Due to resolution requirements and the low signal-to-noise[10] of intact proteins, which necessitates extensive signal averaging, there is generally insufficient instrument time available to select all observed proteoforms for fragmentation on an LC time scale. Additionally, proteoforms selected for fragmentation may not be identified due to low signal-to-noise (S/N) typically observed in MS2 spectra of intact proteins (further exacerbated as mass increases[10]), poor fragmentation or excessively complex fragmentation data.

We have recently developed the open source and freely available software program Proteoform Suite (https://smith-chem-wisc.github.io/ProteoformSuite/), which is able to identify proteoforms in complex data by intact-mass alone.[4, 11–13] Proteoform Suite compares observed proteoform masses to both a database and to co-eluting observed masses, selects frequent mass differences corresponding to modifications, and constructs and visualizes proteoform families from accepted mass differences. Intact-mass analysis and the construction of proteoform families increases the number of proteoform identifications beyond what is identified by top-down alone and enables interesting candidates to be selected for subsequent targeted top-down analysis. In previous work, we have performed intact-mass analysis on isotopically-labeled samples from several biological systems, including yeast[11], *E. coli*[14], and human Jurkat cell lysate[15], and label-free samples from systems of reduced biological complexity including yeast[13] and mouse mitochondria.[12]

Recently, intact-mass analysis was used in combination with targeted MS2 to identify proteoforms >50 kDa in human heart tissue.[16]

Here, we extend this strategy to the 21 Tesla (T) Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometer platform.[17, 18] This mass analyzer offers the highest attainable mass accuracy and resolving power. We used Proteoform Suite to construct proteoform families, combining both top-down results and intact-mass observations in size-separated fractions of cell lysate from the MCF-7 breast cancer cell line analyzed by 21 T FT-ICR MS. 2830 experimental proteoforms were observed, of which 932 were identified, 44 were ambiguous between more than one possible identification, and 1854 were unidentified. Proteoform Suite constructed 520 proteoform families from the data, consisting of 766 unique proteoforms identified by top-down MS2 analysis at 1% FDR using TDPortal and 166 additional proteoforms identified by intact-mass (~4.7% global FDR) using Proteoform Suite. Smith et. al. recently published a proteoform level schema to represent ambiguity in proteoform identifications. Here, we have implemented this proteoform level classification in Proteoform Suite for intact-mass identifications, which provides users with the ambiguity levels and sources of ambiguity for each intact-mass proteoform identification. The results in this study demonstrate that the high-quality data obtained on the 21 T FT-ICR MS platform are well-suited for human proteoform identifications and proteoform family construction by intact-mass analysis.

METHODS

Data acquisition.

Sample preparation—Two pellets of 1×10^7 MCF-7 cells were thawed on ice and resuspended in ten volumes of lysis buffer consisting of 4% SDS (Sigma Aldrich), 100 mM Tris pH 7.5 (TekNova), 10 mM DTT (Sigma Aldrich), 10 mM sodium butyrate (Sigma Aldrich) and 1X Thermo Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific). Cell pellets were lysed by heating at 95 °C for 10 minutes, vortexing every 2 minutes. Cellular debris was removed by centrifugation at 20,000 x g for 20 minutes. Acetone (Sigma Aldrich) protein precipitation was performed on the supernatant and the resulting protein pellet was suspended in 150 μL of 1 % SDS for quantification via bicinchoninic acid (BCA) assay. Size-based separation of approximately 400 μg of each sample into 12 fractions was performed using a GELFrEE 8100 fractionation station with a 10% GELFrEE cartridge (Expedeon) following the manufacturer's recommended procedure. Methanol-chloroform (Honeywell- Burdick & Jackson) extraction was performed on each fraction to remove SDS.[19, 20] Each pellet was reconstituted in solvent A: 0.3% formic acid (Thermo Scientific Pierce), and 5% acetonitrile (Honeywell – Burdick & Jackson) in water with % expressed as v/v).

Liquid Chromatography—For each injection, 2–4 μL of reconstituted intact protein was loaded onto an in-house-fabricated 360 μm O.D. × 150 μm I.D. fused-silica microcapillary trap column packed 2.5 cm with PLRP-S resin (5 μm particle, 1000 Å pore, Agilent Technologies, Palo Alto, CA, USA). The LC system (Acquity M-Class, Waters, Milford, MA, USA) was operated at a flow rate of 2.5 μL/min for loading onto the trap column and

washed with 95% solvent A for 10 min. Separation was achieved on an in-house-fabricated analytical column packed 17.5 cm with the PLRP-S resin. Samples were eluted at a flow rate of 0.3 μL/min over 90 min with the following gradients: (MCF-7 F1&F2) 5–15 %B in 5 min, 15–55 %B in 80 min, 55–75 %B in 5 min; (MCF-7 F3-F6) 5–20 %B in 5 min, 20–60 %B in 80 min, 60–75 %B in 5 min; (MCF-7 F7&F8) 5–25 %B in 5 min, 25–60 %B in 80 min, 60–75 %B in 5 min. The gradients utilized solvent A: 0.3% formic acid and 5% acetonitrile in water, and solvent B: 47.5% acetonitrile, 47.5% 2-propanol, 4.7% water and 0.3% formic acid (% all expressed as v/v). Following separation, proteins were directly ionized by nanoelectrospray ionization (2.5 kV source voltage; 15 V SID) using a 15 μm fused-silica PicoTip emitter (New Objective, Woburn, MA) packed 3 mm with PLRP-S resin.

Mass Spectrometry—The instrument was operated with Xcalibur software (Thermo Fisher Scientific, Waltham, MA, USA), and each fraction injected for 3 separate experiments: two data-dependent CID MS/MS runs and one MS1-only run. All spectra were collected in the ion cyclotron resonance (ICR) mass analyzer at 21 tesla (T). For data-dependent experiments, some data acquisition parameters were varied based upon the expected MW range of the proteins contained within each fraction. For MS1 spectra – resolving power (RP) was set to 300,000 at m/z 400; 1E6 automatic gain control (AGC) target; 4–6 microscans per spectrum; 600–2000 m/z range. For MS2 spectra – RP was set to 150,000 or 300,000 at m/z 400; 5E5 AGC target for CID MS2; 1–2 microscans per spectrum; 300–2000 m/z range. CID activation was performed with 10 m/z isolation width, 35% normalized collision energy, 10 ms activation period, 0.25 q, and 3–6 fragment ion fills of the multipole storage device were performed such that cumulative fragment ion targets were 1.5E6–3.0E6 charges prior to detection in the ICR cell. Data-dependent selection of precursors for MS2 was allowed from 700–1400 m/z, and dynamic exclusion was enabled with a repeat count of one and repeat and exclusion durations set to 240 s. Charge state exclusion was enabled for $[M+H]^{+}$ and $[M+2H]^{2+}$. For all MS1-only experiments (regardless of fraction/MW), RP was set to 300,000 at 400 m/z; 1E6 AGC target; 6 microscans per spectrum; 600–2000 m/z range. All raw data is available on the MassIVE platform with identifier MSV000085978.

Data analysis

Top-Down Data Analysis—The data (.raw files) derived from data-dependent CID MS/MS experiments (2 per fraction for a total of 16 .raw files) were uploaded to the National Resource for Translational and Developmental Proteomics Galaxy[21] web portal for performing top-down proteomics database searches, which is freely available for academic collaborators (http://nrtdp.northwestern.edu/tdportal-request). This platform (TDPortal[22]) utilizes two search modes defined for ProSight PTM 2.0[23, 24]: a narrow absolute mass search (with precursor mass measurement tolerance of 2.2 Da and 10 ppm fragment mass tolerance), and a biomarker search (similar to traditional “no-enzyme” search with biomarker and fragment mass tolerances set to 10 ppm). Details regarding Xtract deconvolution parameters and other aspects of the data analysis can be found within the TDReport file (available on the MassIVE platform with identifier MSV000085978), which can be viewed with TDViewer software (freely available at http://

topdownviewer.northwestern.edu). Top-down hits (proteoform spectrum matches) corresponding to 1% protein-level false discovery rate (FDR) were exported to a Microsoft Excel file.

Deconvolution of MS1-only Files—Proteoform Suite performs intact-mass analysis of proteoform masses in the MS1 spectra. To obtain a list of observed proteoform masses, MS1-only raw files were deconvoluted using Thermo Protein Deconvolution 4.0. We used a fit factor of 70%, minimum S/N of 2, remainder threshold of 10%, minimum detected charge states of 3, and charge range of +5 to +50. A sliding window of 0.5 minutes and 50% offset was used to deconvolute the retention time range of 0 – 100 minutes. A Microsoft Excel file containing the raw experimental components was exported for each raw file.

Mass and Retention Time Calibration in Proteoform Suite—Proteoform Suite version 0.3.6 was used for all analysis (https://github.com/smith-chem-wisc/ProteoformSuite/releases). Mass calibration of deconvolution and TDPortal results was performed in Proteoform Suite as previously described.[13] We implemented a retention time calibration algorithm (described in the Retention Time Calibration section of the Supporting Text) to account for run-to-run variation and the different LC gradients utilized. For both calibrations, well characterized top-down hits ($C\text{-score}^{25} > 40$) were used as calibration points. Isotopic peaks of different charge states were selected using tolerances of ±7 ppm and ±15 minutes in each of the raw data files. These tolerances were selected based on an analysis of top-down precursor mass error and retention time differences (Supporting Figures S-1 and S-2). Calibrated files with top-down hits and raw experimental components were exported and used for subsequent Proteoform Suite analyses.

Proteoform Suite Intact-Mass Analysis—Intact-mass analysis and construction of proteoform families in Proteoform Suite have been previously described.[4, 11–14] Briefly, a theoretical proteoform database was created using a UniProt *Homo sapiens* .xml database downloaded March 2017 containing canonical sequences and Uniprot-annotated modifications and truncations, including signal peptides. Theoretical proteoforms were created using combinations of up to two annotated modifications. The theoretical database contained 58,390 theoretical proteoforms (20,218 unique proteins), 28,832 of which had at least 1 bottom-up peptide (7738 unique proteins).

Raw experimental components were read in from the deconvolution results of the MS1-only raw files, and both monoisotopic mass errors and charge state harmonics were corrected with a 5 ppm tolerance. A list of unique observed experimental proteoforms (intact-mass experimental proteoforms) was created by aggregating the raw experimental components with a mass tolerance of 5 ppm, retention time tolerance of 2.5 min, and a missed monoisotopic mass error of 3 units. Top-down hits were read in, filtered by applying a C-score cutoff of 3, and aggregated with a retention time tolerance of 5 min to generate a list of top-down proteoforms. A theoretical proteoform for each top-down proteoform was added to the theoretical database if not already present. The lists of observed intact-mass experimental proteoforms and top-down proteoforms were combined, removing any intact-mass experimental proteoforms explained by a top-down proteoform (i.e., already identified) utilizing the same tolerances as used for aggregation. The resulting list of experimental

proteoforms contained top-down proteoforms and intact-mass proteoforms corresponding to observed yet proteoforms unidentified by the top-down analysis.

An experimental-theoretical (comparing experimental and theoretical proteoform masses) and experimental-experimental (comparing experimental proteoform masses with one another within 2.5 minute retention time) mass comparisons were performed separately. In each comparison, a delta mass histogram was constructed with a bin size of 0.1 Da (Supporting Figure S-3), and abundant peaks corresponding to known, common modifications were accepted for proteoform family construction. In order to control the size of the database in the experimental-theoretical comparison, we required each theoretical proteoform to be either identified by top-down anlaysis or from a protein identified by bottom-up analysis. The bottom-up analysis was performed with the software program MetaMorpheus[26] on a published MCF-7 dataset[27] (described in the Bottom-Up MetaMorpheus Search section of the Supporting Text).

Proteoform families were constructed from accepted experimental-theoretical and experimental-experimental relations. A mass error tolerance of 1.5 ppm was used for identification to control the FDR. Identification of experimental proteoforms by intact-mass was performed within each proteoform family; beginning with each theoretical proteoform, mass difference connections between proteoforms were followed and experimental proteoforms were assigned an identification until a dead-end was reached or an identification did not meet the mass tolerance (1.5 ppm) or the heuristic criteria (e.g.: loss of acetylation when no acetylation is present on the proteoform). Therefore, intact-mass identifications are made in each proteoform family by first identifying experimental proteoforms in experiment-theoretical pairs in each family, and subsequent experiment-experiment connections are used to identify connected experimental proteoforms. Each experiment-experiment pair consists of two experimental proteoforms that are within the 2.5 min retention time tolerance set during the experiment-experiment delta mass comparison. If an intact-mass identification was ambiguous between a top-down identification and another possible identification, the top-down identification was utilized. Relations between proteoforms that did not result in an identification were removed, and proteoform families were reconstructed from accepted relations. Identifications were exported in a tab-delimited text file, and redundant identifications were manually removed. Decoy proteoform families were constructed as previously described[11, 13], and a global false discovery rate was calculated by dividing the average number of proteoforms identified in decoy families by the number of proteoforms identified in target families.

Proteoform Suite Top-Down Hit Precursor Analysis—We evaluated how intact-mass analysis performed on top-down validated identifications. We created a tab-delimited text file from unique top-down identified proteoforms that had a minimum C-score of 40 (well-characterized proteoforms) and input this file into Proteoform Suite. We performed an intact-mass analysis utilizing the same aggregation parameters and database described above. We compared the Proteoform Suite intact-mass identifications to the top-down identifications determined by TDPortal in Microsoft Excel.

RESULTS AND DISCUSSION

Proteoforms and Proteoform Families

From analysis of eight GELFrEE fractions analyzed by 21 T FT-ICR MS, the 16 top-down data-dependent CID MS/MS .raw files were searched against a database of candidate human proteoforms with TDPortal. This search resulted in the identification of 354 unique proteins (defined by UniProt accession numbers) expressed as 1,684 unique proteoforms (defined by Proteoform Record, PFR; Consortium for Top-Down Proteomics Proteoform Repository http://repository.topdownproteomics.org/) at 1% FDR. Of these 1,684 proteoforms identified by TDPortal, 766 proteoforms from 339 unique proteins exhibited a C-score of 3 or greater (Supporting Table S-1), which was used as a threshold when importing top-down results into Proteoform Suite for intact-mass analysis. This is considered the minimum C-score for a proteoform to be identified; proteoforms with C-score > 40 are considered both identified and well-characterized.[25]

The MS1-only raw files were deconvoluted, revealing an additional 2064 intact-mass experimental proteoforms observed but not identified by top-down analysis, of which 166 were identified at a calculated global FDR of 4.7% (Supporting Table S-2), 44 were ambiguous (Supporting Table S-3), and 1854 were unidentified (Supporting Table S-4) by Proteoform Suite. The intact-mass FDR is calculated by constructing decoy proteoform families from decoy experiment-theoretical and experiment-experiment relations, and determining the ratio of decoy intact-mass identifications made in the decoy families to target identifications made in the target families, as described in detail previously.[11, 13] There were an additional 43 protein accessions identified by intact-mass analysis. In the experimental-theoretical comparison, any intact-mass match with a theoretical proteoform required either a top-down proteoform ID or at least one bottom-up peptide identification to prevent false intact-mass identifications. A summary of the results is shown in Figure 1.

Proteoform Suite constructs proteoform families and provides visualization as a network of nodes (unique proteoform masses) and edges (mass differences between proteoforms). The visualized 520 proteoform families are shown in Figure 2A. There were 336 identified proteoform families (1 gene), 14 ambiguous families (more than one gene) and 170 unidentified families (no gene). We selected several examples that exemplify how intact-mass analysis complements top-down analysis. Proteoform Suite identified two 35 kDa proteoforms (acetylated and phosphorylated, and acetylated) from the YBX1 Y-Box binding protein gene, which were not identified by top-down analysis (Figure 2B). Larger proteoforms are particularly difficult to identify by top-down analysis because of the lower S/N inherent to larger mass and other factors.[10] Top-down analysis identified an unmodified proteoform from the gene NDUFC2. Proteoform Suite was able to identify an acetylated proteoform from this gene by intact-mass experimental-experimental comparison (Figure 2C). Proteoform Suite could not have identified this family by intact-mass alone because the acetylation is not annotated in the database (+42 Da) and the methionine was not cleaved (+131 Da). As a result, the mass difference between the observed experimental proteoform and the theoretical proteoform in the database was +173 Da, which was not an accepted mass difference in the experimental-theoretical comparison. This example illustrates how

MS2 identification was necessary for initial identification, and Proteoform Suite was able to leverage this identification to acquire additional IDs.

Many of the intact-mass identifications were for modified proteoforms. Of the 188 new intact-mass identifications, there were 14 unmodified intact-mass identifications, 114 with at least one acetylation, 27 with at least one phosphorylation, and 62 with at least one methylation. We observed 82 intact-mass proteoforms exhibiting a mass shift of 98.06 Da, which could potentially be an acetone adduct from the acetone precipitation performed.[28] There were 77 intact-mass proteoforms exhibiting a mass shift of 266.15, which corresponds to an SDS adduct, and 343 proteoforms with at least one oxidation. Intact-mass identifications resulting from an oxidation, 98.06 Da shift, or SDS adducts were not counted as additional identifications in number reporting (Figure 1). Although these modifications are likely sample handling artifacts, it is still important to identify these species to prevent misidentification and to potentially include them in quantitative analyses. Due to false discovery constraints, intact-mass analysis is limited to a theoretical database with canonical sequences and a small number of annotated PTMs (in this study, combinations of up to two PTMs). However, the experiment-experiment comparison enables heavily modified proteoforms to be identified; if a proteoform with fewer modifications is identified in the experiment-theoretical comparison or through top-down MS/MS analysis, proteoforms from the same family with additional modifications can be identified through the experiment-experiment delta mass comparison.

At this time, identification of proteoforms by intact-mass alone in Proteoform Suite is also limited to common modifications that are observed at high frequency in the experimental-experimental mass comparison. If an uncommon modification is expected on a proteoform, adding this modification to the database could enable its identification by Proteoform Suite. Intact-mass analysis alone cannot localize modifications, so dynamic modification sites cannot be revealed by intact-mass analysis alone. However, knowledge of proteoform identity gives researchers the opportunity to better gauge whether subsequent targeted top-down experiments should be performed to localize modifications and confirm identifications returned by Proteoform Suite.

Proteoform Level Classification

Smith et. al recently introduced the five-level proteoform classification system, which indicates the amount of ambiguity in a given proteoform identification, including ambiguity in gene of origin, modification identification, modification localization, or sequence.[29] Level 1 proteoform identifications have no sources of ambiguity, Level 2’s have one source of ambiguity, Level 3’s have two different sources of ambiguity, Level 4’s have three sources of ambiguity, and a Level 5 indicates that no information other than the observed mass of the proteoform is known. As described above, there were 166 intact-mass identifications and 44 ambiguous identifications. The intact-mass identifications do not have ambiguity in modification identification or sequence, whereas the ambiguous identifications have ambiguity in any of the four possible sources listed.

We determined the proteoform level for each of the 166 intact-mass identifications, of which 13 were Level 1, 152 were Level 2, and 1 was Level 3. Of the 152 Level 2 identifications,

151 were ambiguous with respect to PTM localization and 1 was ambiguous with respect to gene of origin. The 1 Level 3 identification was ambiguous due to both PTM localization and gene of origin. Of the 44 ambiguous intact-mass experimental proteoforms (Supporting Table S-3), 11 were Level 3, 3 were Level 4, and 30 were Level 5. Although these proteoforms are ambiguous, the provided candidates can be helpful when searching for proteins of interest and a subsequent targeted top-down analysis could determine their identity. The 1854 unidentified experimental proteoforms were also Level 5 assignments (Supporting Table S-4).

Modifications are not localized in intact-mass analysis, so any intact-mass identifications of a modified proteoforms must be assigned as Level 2 or higher due to ambiguity in PTM localization. The level of ambiguity is dependent on the size of the theoretical database and search parameters utilized; for example, the number of PTM combinations allowed and the mass tolerances utilized could all affect how many theoretical proteoforms match each experimental mass. However, it is still useful to know for each intact-mass identification the sources and levels of ambiguity within the context of the search space and database utilized. Future analyses could integrate bottom-up modified peptide assignments, modification annotations in repositories, or subsequent targeted top-down analyses to localize PTMs and reduce the ambiguity of such identifications.

Top-Down Precursor Intact-Mass Analysis

We compared our 21 T FT-ICR top-down results with a previously acquired QE-HF Orbitrap top-down dataset of GELFrEE fractionated MCF-7 cell lysate (data available on the MassIVE platform with identifier MSV000086148). The samples were separated on a 10% GELFrEE cartridge, and fractions 1 through 6 were analyzed (one MS/MS technical replicate). The LC-MS parameters are described in the QE-HF Data Acquisition section of the Supporting Text. We compared the 1% protein-level FDR results from the QE-HF TDPortal MS/MS search (Supporting Table S-6) to the results from fractions 1 through 6 (first technical replicate each) of the 21 T FT-ICR TDPortal search to determine the mass accuracy improvements provided by the 21 T FT-ICR platform for human proteoform analysis.

A mass error histogram for all top-down proteoform hits is shown in Figure 3, indicating that the 21 T platform yielded increased mass accuracy for proteoform precursor masses. Missed monoisotopic mass errors were corrected prior to calculating the mass error. Approximately 47% of the 21 T FT-ICR top-down hits had sub-ppm precursor mass accuracy, whereas approximately 20% of the QE-HF top-down hits had sub-ppm precursor mass accuracy. It is important to note that mass error depends significantly on the averagine fits used to determine the monoisotopic mass because mass accuracy is limited to the difference between the averagine and true elemental composition of the analyte.[30] For example, in an analysis of a monoclonal antibody at 21 T, manual examination of light chain fragment monoisotopic masses yielded 0.3 ppm RMS mass error based on elemental composition. Following monoisotopic mass assignments via average fitting, RMS errors of 3.3 ppm were observed for the same data.[31] This issue, combined with difficulties associated with missed monoisotopic mass assignments, was the primary motivation behind the

inclusion of an isotope filter algorithm in TDValidator software (Proteinaceous, Inc. Evanston, IL) that can match experimental isotopic peak clusters from original raw data with calculated isotopic clusters given a specific sequence. Expanding the use of these algorithms to high-throughput top-down proteomics experiments would enable proteoform analysis to more effectively realize the benefits of state-of-the-art high-resolution mass spectrometers like the 21 T.

We performed a separate Proteoform Suite analysis on top-down precursor masses from TDPortal to evaluate the performance of intact-mass identifications returned by Proteoform Suite. We created a deconvolution result input file for Proteoform Suite using the precursor mass, a constant intensity (this value is not reported by TDPortal), and retention time for each top-down proteoform that had been identified and well characterized by TDPortal (minimum C-score 40). We used the same theoretical database and Proteoform Suite parameters comparable to those utilized for the intact-mass analysis described in the Methods. Of the 489 top-down proteoforms which met the minimum C-score threshold, 237 were identified in Proteoform Suite, 6 were ambiguous, and 246 were unidentified (Supporting Table S-6). For the identified proteoforms, we compared the accession numbers, the unlocalized modifications, and the sequence to the original list of top-down proteoforms to confirm the matches. 221 of the top-down proteoforms matched, and 16 did not (6.8% FDR). However, we noticed that many of the proteoforms that did not match a top-down hit were histone proteoforms, which have high sequence homology and are heavily modified, so could easily match a different histone identification. When considering only non-histone proteoforms, 216 matched the top-down proteoform identification, and 10 did not, resulting in a calculated FDR of 4.4%, which is close to the FDR or 4.7% determined in the intact-mass analysis described above. There were 2 ambiguous proteoforms and 201 unidentified proteoforms in the non-histone intact-mass analysis. There was a corresponding theoretical proteoform in the theoretical database for each top-down identification; therefore, the unidentified proteoforms were unidentified due to either a missed monoisotopic mass error in the deconvolution step or large precursor mass error.

One difference between a typical intact-mass analysis and this top-down precursor intact-mass analysis is that using such a file of true precursor masses means that every mass on the list is a likely real proteoform feature, and therefore the list does not include deconvolution artifacts (apart from missed monoisotopic errors). Deconvolution artifacts are a continuing challenge in top-down proteomics and even more so in intact-mass analysis, which relies heavily on the quality of the deconvolution output.

A challenging remaining problem is how to evaluate the fidelity of family construction for unidentified proteoform families. One possibility for future analysis of unidentified proteoform families is to perform a targeted top-down analysis of intact-mass proteoforms observed in unidentified proteoform families. Once the proteoforms in a family are confidently identified, the mass differences between proteoforms generated by Proteoform Suite could be evaluated for accuracy, i.e., to evaluate whether the proteoforms are grouped into the correct proteoform families.

As discussed in the Methods, the Proteoform Suite theoretical database was created using sequences from a human canonical database downloaded from UniProt with PTM combinations of up to two annotated PTMs. A theoretical proteoform for each top-down proteoform was then added to the theoretical database if not already present. Of the 489 top-down proteoforms that met the C-score threshold, 277 did not have a corresponding theoretical proteoform mass in the database before supplementation with top-down identifications. This can be due to a larger number of modifications, an amino acid variant, or an unannotated truncation event. The significant proportion of MS2 identified proteoforms that would not been included as a theoretical proteoform in the Proteoform Suite database without a prior top-down analysis shows why intact-mass analysis is highly improved when integrated with top-down analysis; many proteoforms cannot be identified without MS2 data. Intact-mass analysis is however able to leverage these top-down MS2 identifications to identify additional, co-eluting proteoforms from the same proteoform family. As proteoform identifications are continuously catalogued, databases will be more customized to a given sample and thereby increase the number of intact-mass identifiable proteoforms. Intact-mass analysis offers a simple approach to identify MS2-identified proteoforms in subsequent analyses, which will be particularly useful in quantitative and biological studies. This is one powerful motivation for the construction of deep proteoform catalogs for widely used cell lines, which will facilitate the rapid MS1-based identification and quantification of proteoforms

Unidentified Proteoform Families

One unique attribute of Proteoform Suite is that proteoform families are constructed and visualized even for families without an identification. Of the 1854 unidentified experimental proteoforms, 463 were in the 170 unidentified families, and 1391 were orphans (no accepted experimental-theoretical or experimental-experimental relations formed). The orphan experimental proteoforms present a greater challenge to identification because no information is known about these experimental proteoforms other than their mass and retention time; additionally, it is difficult to know without manual inspection whether these orphan experimental proteoforms are deconvolution artifacts or real proteoforms observed in the MS1 spectra. A tradeoff in deconvolution is using criteria to prevent as many artifacts as possible without filtering out low abundance experimental proteoform observations.

Fortunately, the unidentified families present additional information about the experimental proteoforms given the observed mass differences. For example, a family with multiple methylations was observed but remains unidentified (Figure 2D). Construction of families enables organization of observed intact-masses, which facilitates selection of interesting candidates for subsequent targeted analysis, such as the family in Figure 2D. A potential further development of proteoform family analysis could be network analysis to determine whether an experimental or biological condition increases the numbers of certain types of modifications, such as phosphorylation, or of certain types of artifact, such as SDS adduction.

As discussed above, many experimental proteoforms may be unidentified due to the theoretical database utilized, which only contained canonical sequences, combinations of up

to two annotated PTMs, and top-down identified proteoforms. Alternative splicing, amino acid variants, heavily modified proteoforms, and proteoforms with novel modifications all present major challenges for intact-mass analysis. Integration of different types of data, including data from genomic sequencing and bottom-up analyses, could also facilitate identification of more experimental proteoforms by intact-mass.[14]

As expected[18], the number of proteoforms identified decreased as a function of molecular weight (MW). This is largely due to decreased sensitivity inherent to mass spectrometric analysis (on LC-timescales) of high MW proteins electrosprayed under denaturing conditions, as well as the need for improved separation of larger proteins by reversed phase LC.[9, 10] For these experiments, a delta mass mode (wide absolute mass search) was not performed in TDPortal, which is particularly detrimental to the identification of high-MW proteoforms. Additionally, all MS/MS spectra were taken as a single or the sum of two transients. Sampling rate was prioritized over spectral quality (higher quality is achieved via additional signal averaging, requiring additional time per scan) to the detriment of high-MW proteoform identifications. Despite this, over 193 proteins expressed as 428 proteoforms were identified in MCF-7 fractions 7 and 8 at 1% protein-level FDR. These proteoforms ranged in size from 4–48 kDa; 34 MCF-7 proteoforms >30 kDa were identified in total by TDPortal.

Increases in the number of proteoform identifications in top-down searches will improve intact-mass analysis. Top-down identification of just one proteoform in a previously unidentified family enables other proteoforms in the family to be identified by intact-mass analysis. One strategy for increasing the number of proteoform identifications in MS/MS search software programs is implementation of an open mass search, where the precursor mass tolerance is widened to enable identification of proteoforms containing unexpected PTMs not annotated in the database.[23, 24, 32, 33] A remaining challenge is interpreting these large delta mass differences between the observed proteoform precursor mass and the match in the theoretical proteoform database. This mass shift interpretation will be necessary for integration of such top-down results in Proteoform Suite. One possibility is to employ the two-pass search strategy Global-PTM-Discovery (G-PTM-D)[34], implemented in MetaMorpheus.[26] A first pass open precursor mass search selects for discovered PTMs and adds these to the database; a second search is performed, localizing PTMs and reporting identifications with the novel PTMs localized. Future work will optimize this strategy for top-down proteoform analysis in MetaMorpheus to identify novel PTM-containing proteoforms that were selected for MS2. These additional identifications will further improve proteoform family identification and intact-mass analysis by increasing the number of identified observed proteoform masses. As more comprehensive top-down analyses are performed and proteoforms are documented in repositories such as the Proteoform Atlas maintained by the Consortium for Top-Down Proteomics, identification by intact-mass will be increasingly useful for identifying larger numbers of proteoforms from complex samples.

CONCLUSIONS

We used the freely available and open-source software program Proteoform Suite to construct human proteoform families from MCF-7 data acquired on the 21 T FT-ICR MS.

TDPortal identified 1,694 unique proteoforms at 1% FDR, 766 of which had a C-score of 3 or greater. From these 766 top-down proteoforms and the 2830 intact-mass experimental proteoforms observed, we constructed 520 proteoform families with Proteoform Suite. 166 additional proteoforms were identified by intact-mass (~4.7% global FDR) using Proteoform Suite. The remaining 1854 were unidentified observed proteoforms, 463 of which contained at least one proteoform relation to another co-eluting experimental proteoform. We performed an intact-mass analysis in Proteoform Suite of precursor masses confidently identified by MS2 in TDPortal and found that when histones were excluded from the analysis results, there was a false identification rate of 4.4%. This analysis shows how the 21T FT-ICR MS platform enables intact-mass identifications in complex biological samples. Construction of proteoform families and intact-mass analysis offer a way to identify additional proteoforms in top-down analyses, visualize results, and provide interesting targeted top-down candidates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R35GM126914. A portion of this work was performed at the Ion Cyclotron Resonance User Facility at the National High Magnetic Field Laboratory, which is supported by the National Science Foundation Division of Materials Research and Division of Chemistry through DMR 1644779, DMR 1157490, and the State of Florida. L.V.S. was supported by the NIH Biotechnology Training Program, T32GM008349. R.M.M. was supported by the NIH Chemistry Biology Interface Training program, T32GM008505.

REFERENCES

1. Aebersold R; Agar JN; Amster IJ; Baker MS; Bertozzi CR; Boja ES; Costello CE; Cravatt BF; Fenselau C; Garcia BA; Ge Y; Gunawardena J; Hendrickson RC; Hergenrother PJ; Huber CG; Ivanov AR; Jensen ON; Jewett MC; Kelleher NL; Kiessling LL; Krogan NJ; Larsen MR; Loo JA; Ogorzalek Loo RR; Lundberg E; MacCoss MJ; Mallick P; Mootha VK; Mrksich M; Muir TW; Patrie SM; Pesavento JJ; Pitteri SJ; Rodriguez H; Saghatelian A; Sandoval W; Schluter H; Sechi S; Slavoff SA; Smith LM; Snyder MP; Thomas PM; Uhlen M; Van Eyk JE; Vidal M; Walt DR; White FM; Williams ER; Wohlschlager T; Wysocki VH; Yates NA; Young NL; Zhang B, How many human proteoforms are there? Nat Chem Biol 2018, 14 (3), 206–214. [PubMed: 29443976]
2. Yang X; Coulombe-Huntington J; Kang S; Sheynkman GM; Hao T; Richardson A; Sun S; Yang F; Shen YA; Murray RR; Spirohn K; Begg BE; Duran-Frigola M; MacWilliams A; Pevzner SJ; Zhong Q; Trigg SA; Tam S; Ghamsari L; Sahni N; Yi S; Rodriguez MD; Balcha D; Tan G; Costanzo M; Andrews B; Boone C; Zhou XJ; Salehi-Ashtiani K; Charloteaux B; Chen AA; Calderwood MA; Aloy P; Roth FP; Hill DE; Iakoucheva LM; Xia Y; Vidal M, Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell 2016, 164 (4), 805–17. [PubMed: 26871637]
3. Smith LM; Kelleher NL; Consortium for Top Down Proteomics, Proteoform: a single term describing protein complexity. Nat Methods 2013, 10 (3), 186–7. [PubMed: 23443629]
4. Shortreed MR; Frey BL; Scalf M; Knoener RA; Cesnik AJ; Smith LM, Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. J Proteome Res 2016, 15 (4), 1213–21. [PubMed: 26941048]
5. Catherman AD; Skinner OS; Kelleher NL, Top Down proteomics: facts and perspectives. Biochem Biophys Res Commun 2014, 445 (4), 683–93. [PubMed: 24556311]
6. Schaffer LV; Millikin RJ; Miller RM; Anderson LC; Fellers RT; Ge Y; Kelleher NL; LeDuc RD; Liu X; Payne SH; Sun L; Thomas PM; Tucholski T; Wang Z; Wu S; Wu Z; Yu D; Shortreed MR; Smith

LM, Identification and Quantification of Proteoforms by Mass Spectrometry. Proteomics 2019, 19, e1800361. [PubMed: 31050378]

7. McCool EN; Lubeckyj RA; Shen X; Chen D; Kou Q; Liu X; Sun L, Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. Anal Chem 2018, 90 (9), 5529–5533. [PubMed: 29620868]
8. Zhao Y; Sun L; Zhu G; Dovichi NJ, Coupling Capillary Zone Electrophoresis to a Q Exactive HF Mass Spectrometer for Top-down Proteomics: 580 Proteoform Identifications from Yeast. J Proteome Res 2016, 15 (10), 3679–3685. [PubMed: 27490796]
9. Cai W; Tucholski T; Chen B; Alpert AJ; McIlwain S; Kohmoto T; Jin S; Ge Y, Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. Anal Chem 2017, 89 (10), 5467–5475. [PubMed: 28406609]
10. Compton PD; Zamdborg L; Thomas PM; Kelleher NL, On the scalability and requirements of whole protein mass spectrometry. Anal Chem 2011, 83 (17), 6868–74. [PubMed: 21744800]
11. Cesnik AJ; Shortreed MR; Schaffer LV; Knoener RA; Frey BL; Scalf M; Solntsev SK; Dai Y; Gasch AP; Smith LM, Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. J Proteome Res 2018, 17 (1), 568–578. [PubMed: 29195273]
12. Schaffer LV; Rensvold JW; Shortreed MR; Cesnik AJ; Jochem A; Scalf M; Frey BL; Pagliarini DJ; Smith LM, Identification and Quantification of Murine Mitochondrial Proteoforms Using an Integrated Top-Down and Intact-Mass Strategy. J Proteome Res 2018, 17 (10), 3526–3536. [PubMed: 30180576]
13. Schaffer LV; Shortreed MR; Cesnik AJ; Frey BL; Solntsev SK; Scalf M; Smith LM, Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. Anal Chem 2018, 90 (2), 1325–1333. [PubMed: 29227670]
14. Dai Y; Shortreed MR; Scalf M; Frey BL; Cesnik AJ; Solntsev S; Schaffer LV; Smith LM, Elucidating Escherichia coli Proteoform Families Using Intact-Mass Proteomics and a Global PTM Discovery Database. J Proteome Res 2017, 16 (11), 4156–4165. [PubMed: 28968100]
15. Dai Y; Buxton KE; Schaffer LV; Miller RM; Millikin RJ; Scalf M; Frey BL; Shortreed MR; Smith LM, Constructing Human Proteoform Families Using Intact-Mass and Top-Down Proteomics with a Multi-Protease Global Post-Translational Modification Discovery Database. J Proteome Res 2019, 18 (10), 3671–3680. [PubMed: 31479276]
16. Schaffer LV; Tucholski T; Shortreed MR; Ge Y; Smith LM, Intact-Mass Analysis Facilitating the Identification of Large Human Heart Proteoforms. Anal Chem 2019, 91 (17), 10937–10942. [PubMed: 31393705]
17. Hendrickson CL; Quinn JP; Kaiser NK; Smith DF; Blakney GT; Chen T; Marshall AG; Weisbrod CR; Beu SC, 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer: A National Resource for Ultrahigh Resolution Mass Analysis. J Am Soc Mass Spectrom 2015, 26 (9), 1626–32. [PubMed: 26091892]
18. Anderson LC; DeHart CJ; Kaiser NK; Fellers RT; Smith DF; Greer JB; LeDuc RD; Blakney GT; Thomas PM; Kelleher NL; Hendrickson CL, Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. J Proteome Res 2017, 16 (2), 1087–1096. [PubMed: 27936753]
19. Donnelly DP; Rawlins CM; DeHart CJ; Fornelli L; Schachner LF; Lin Z; Lippens JL; Aluri KC; Sarin R; Chen B; Lantz C; Jung W; Johnson KR; Koller A; Wolff JJ; Campuzano IDG; Auclair JR; Ivanov AR; Whitelegge JP; Pasa-Tolic L; Chamot-Rooke J; Danis PO; Smith LM; Tsybin YO; Loo JA; Ge Y; Kelleher NL; Agar JN, Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 2019, 16 (7), 587–594. [PubMed: 31249407]
20. Wessel D; Flugge UI, A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 1984, 138 (1), 141–3. [PubMed: 6731838]
21. Afgan E; Baker D; Batut B; van den Beek M; Bouvier D; Cech M; Chilton J; Clements D; Coraor N; Gruning BA; Guerler A; Hillman-Jackson J; Hiltemann S; Jalili V; Rasche H; Soranzo N; Goecks J; Taylor J; Nekrutenko A; Blankenberg D, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 2018, 46 (W1), W537–W544. [PubMed: 29790989]

22. Toby TK; Fornelli L; Srzentic K; DeHart CJ; Levitsky J; Friedewald J; Kelleher NL, A comprehensive pipeline for translational top-down proteomics from a single blood draw. Nat Protoc 2019, 14 (1), 119–152. [PubMed: 30518910]

23. LeDuc RD; Taylor GK; Kim YB; Januszyk TE; Bynum LH; Sola JV; Garavelli JS; Kelleher NL, ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res 2004, 32 (Web Server issue), W340–5. [PubMed: 15215407]

24. Zamdborg L; LeDuc RD; Glowacz KJ; Kim YB; Viswanathan V; Spaulding IT; Early BP; Bluhm EJ; Babai S; Kelleher NL, ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res 2007, 35 (Web Server issue), W701–6. [PubMed: 17586823]

25. LeDuc RD; Fellers RT; Early BP; Greer JB; Thomas PM; Kelleher NL, The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. J Proteome Res 2014, 13 (7), 3231–40. [PubMed: 24922115]

26. Solntsev SK; Shortreed MR; Frey BL; Smith LM, Enhanced Global Post-translational Modification Discovery with MetaMorpheus. J Proteome Res 2018, 17 (5), 1844–1851. [PubMed: 29578715]

27. Geiger T; Wehner A; Schaab C; Cox J; Mann M, Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics 2012, 11 (3), M111 014050.

28. Guray MZ; Zheng S; Doucette AA, Mass Spectrometry of Intact Proteins Reveals +98 u Chemical Artifacts Following Precipitation in Acetone. J Proteome Res 2017, 16 (2), 889–897. [PubMed: 28088865]

29. Smith LM; Thomas PM; Shortreed MR; Schaffer LV; Fellers RT; LeDuc RD; Tucholski T; Ge Y; Agar JN; Anderson LC; Chamot-Rooke J; Gault J; Loo JA; Pasa-Tolic L; Robinson CV; Schluter H; Tsybin YO; Vilaseca M; Vizcaino JA; Danis PO; Kelleher NL, A five-level classification system for proteoform identifications. Nat Methods 2019, 16 (10), 939–940. [PubMed: 31451767]

30. Senko MW; Beu SC; McLaffertycor FW, Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J Am Soc Mass Spectrom 1995, 6 (4), 229–33. [PubMed: 24214167]

31. Fornelli L; Srzentic K; Huguet R; Mullen C; Sharma S; Zabrouskov V; Fellers RT; Durbin KR; Compton PD; Kelleher NL, Accurate Sequence Analysis of a Monoclonal Antibody by Top-Down and Middle-Down Orbitrap Mass Spectrometry Applying Multiple Ion Activation Techniques. Anal Chem 2018, 90 (14), 8421–8429. [PubMed: 29894161]

32. Park J; Piehowski PD; Wilkins C; Zhou M; Mendoza J; Fujimoto GM; Gibbons BC; Shaw JB; Shen Y; Shukla AK; Moore RJ; Liu T; Petyuk VA; Tolic N; Pasa-Tolic L; Smith RD; Payne SH; Kim S, Informed-Proteomics: open-source software package for top-down proteomics. Nat Methods 2017, 14 (9), 909–914. [PubMed: 28783154]

33. Kou Q; Xun L; Liu X, TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 2016, 32 (22), 3495–3497. [PubMed: 27423895]

34. Li Q; Shortreed MR; Wenger CD; Frey BL; Schaffer LV; Scalf M; Smith LM, Global Post-Translational Modification Discovery. J Proteome Res 2017, 16 (4), 1383–1390. [PubMed: 28248113]

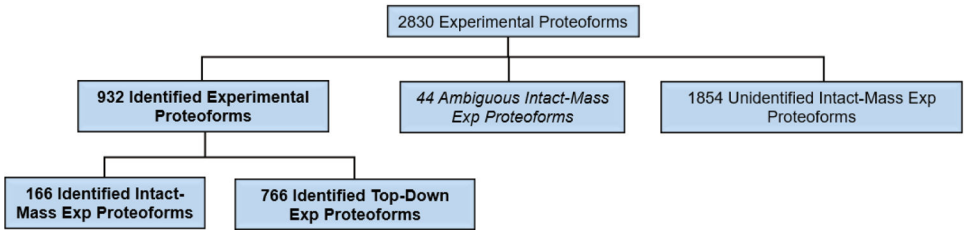


Figure 1.

Summary of top-down and intact-mass results from analysis of MCF-7 cell lysate using 21 T FT-ICR mass spectrometry.

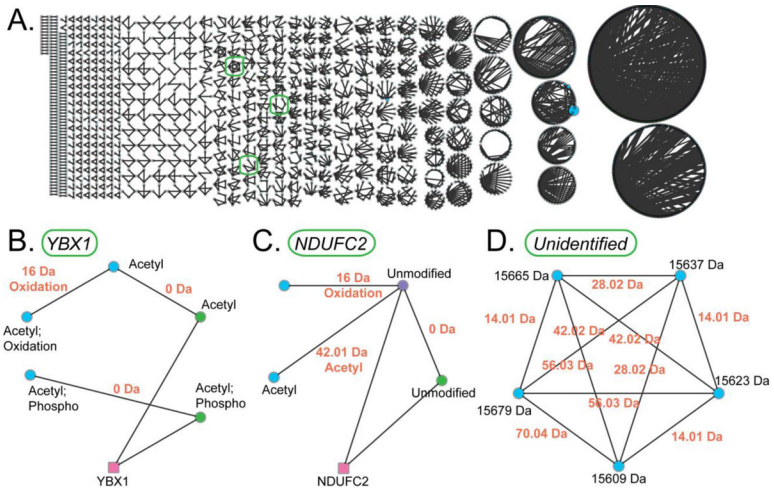


Figure 2.

A) All 542 visualized proteoform families from MCF-7 human lysate. Green squares indicate the proteoform families that are expanded in **2B, 2C**, and **2D**. In these visualized proteoform families, nodes represent proteoform masses and edges represent mass differences corresponding to a modification (0 Da for exact match). Blue circles are intact-mass experimental proteoforms, purple circles are top-down experimental proteoforms, and green circles are theoretical proteoforms from the database. Pink squares represent genes. **B)** Visualized proteoform family from YBX1 gene; intact-mass analysis identified three 35 kDa proteoforms. **C)** Visualized proteoform family from the NDUFC2 gene. Top-down MS2 analysis identified the unmodified proteoform, and intact-mass analysis leveraged this identification to identify two additional modified proteoforms. **D).** An unidentified proteoform family with multiple methylations present.

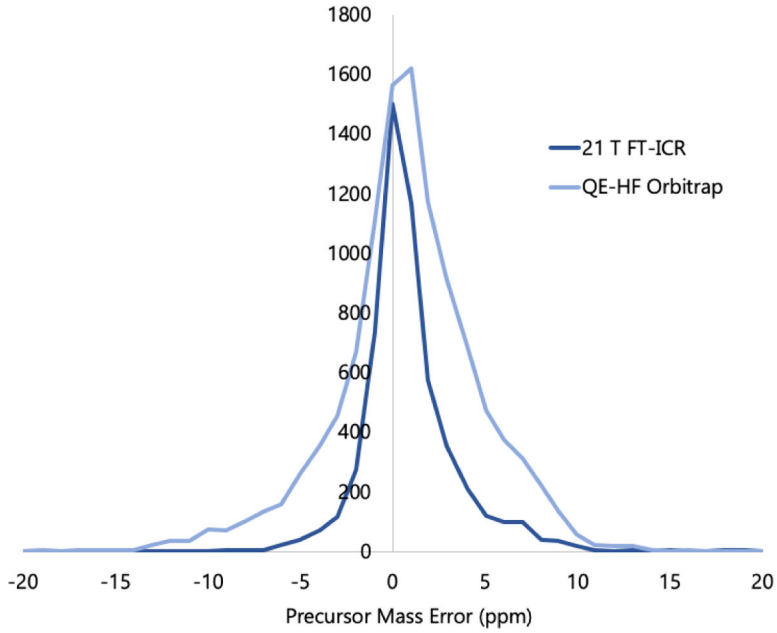


Figure 3.

Histogram of precursor mass error for all top-down hits for 21 T FT-ICR and QE-HF Orbitrap top-down analysis of six fractions (1 technical replicate each) of MCF-7 cell lysate.