Journal of proteome research

Article

pubs.acs.org/jpr

Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements

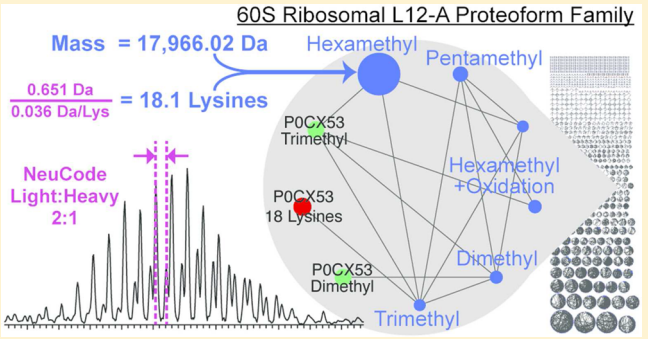
Michael R. Shortreed,[†] Brian L. Frey,[†] Mark Scalf,[†] Rachel A. Knoener,[†] Anthony J. Cesnik,[†] and Lloyd M. Smith*[,†,‡]

[†]Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, Wisconsin 53706, United States

[‡]Genome Center of Wisconsin, University of Wisconsin, 425G Henry Mall, Room 3420, Madison, Wisconsin 53706, United States

Supporting Information

ABSTRACT: Proteomics is presently dominated by the "bottom-up" strategy, in which proteins are enzymatically digested into peptides for mass spectrometric identification. Although this approach is highly effective at identifying large numbers of proteins present in complex samples, the digestion into peptides renders it impossible to identify the proteoforms from which they were derived. We present here a powerful new strategy for the identification of proteoforms and the elucidation of proteoform families (groups of related proteoforms) from the experimental determination of the accurate proteoform mass and number of lysine residues contained. Accurate proteoform masses are determined by standard LC−MS analysis of undigested protein mixtures in an Orbitrap mass spectrometer, and the lysine count is determined using the NeuCode isotopic tagging method. We demonstrate the approach in analysis of the yeast proteome, revealing 8637 unique proteoforms and 1178 proteoform families. The elucidation of proteoforms and proteoform families afforded here provides an unprecedented new perspective upon proteome complexity and dynamics.



KEYWORDS: *proteoform, proteoform family, top-down, proteomics, PTM, database search, NeuCode*

INTRODUCTION

The dominant means for identification of proteins in complex mixtures is bottom-up proteomics.[1] In this approach, a mixture of proteins from the sample of interest is cleaved into peptides, typically using trypsin, and analyzed by liquid chromatography−mass spectrometry (LC−MS). Fragmentation of the peptides within the mass spectrometer yields product-ion mass spectra, which are compared to theoretical mass spectra produced in silico based upon a generic reference protein database of the organism under study. Statistical analysis of the results provides a list of peptides identified in the sample, subject to a specified false discovery rate (FDR).[2] Proteins present in the sample are then inferred from the identified peptides in a process referred to as protein inference.[3,4] Implementations of this approach are routinely able to identify thousands of proteins in yeast,[5] human,[6] or other organisms.[7] The strategy can reveal differences in protein expression in different cell types or in response to cellular growth conditions or treatment with drugs.[8]

While the bottom-up strategy is powerful and widely practiced, it does suffer from major shortcomings. Proteins produced from the same gene can vary substantially in their molecular structure: genetic variations, splice variants, RNA edits, and post-translational modifications (PTMs) all give rise to different forms of the proteins, referred to as "proteoforms".[9] Knowledge of the proteoforms that are present in a system under study is absolutely essential to understanding that system, as the different proteoforms often have dramatically different functional behavior,[10] and regulation of their production is a central aspect of pathway control. One recent example is the finding that intact and clipped human histones differ in post-translational modification patterns[11] and that these combinations of sequence-length and PTM differences have functional consequences. Bottom-up strategies are unable to identify proteoforms for two reasons: first, the digestion of the proteins into peptides means that information is lost as to the protein context within which that peptide is found, making impossible the identification of the parent proteoform from which each peptide is derived; and second, the databases used for peptide identification do not generally contain information regarding amino-acid variant or modified peptides, causing such peptides to be effectively invisible in the absence of specialized search strategies,[12−14] which can introduce problems with search time and false identifications.

Received: November 30, 2015
Published: March 4, 2016

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

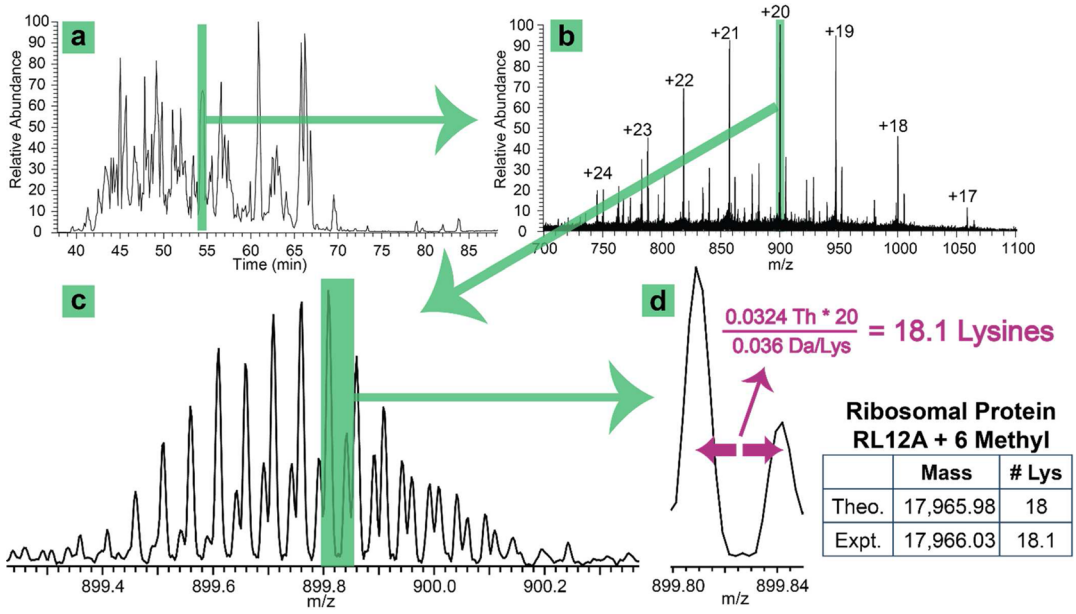


Figure 1. Example intact protein chromatogram and spectrum. (A) An LC−MS chromatogram for one gel electrophoresis fraction of NeuCode SILAC yeast. (B) A full-scan mass spectrum obtained at a resolution of 100 000. (C) An expanded view of the mass spectrum showing one charge-state envelope containing multiple isotope peaks for each of the two isotopologues. (D) A further expanded view displaying the "Light" (left peak) and "Heavy" (right peak) isotopologues; the spacing between these two peaks is used to determine the number of lysines in this proteoform.

One way in which these issues have been addressed is through the alternative strategy of "top-down" proteomics.[15] In top-down proteomics, the entire intact protein is subjected to fragmentation, and the proteoform is identified from the parent mass and fragmentation products; in the ideal case, this can yield a precise identification of the proteoform, including the nature and positions of PTMs. While historically such efforts have largely been limited to the exhaustive study of individual purified protein species,[16] recent work has extended the approach to highly complex samples such as yeast[17] and human[18] cell lysates. However, the highly complex nature and voluminous quantities of the data produced, as well as the need for long MS analysis times to produce data of sufficient sensitivity and resolution for proteoform identification, make this approach, at present, a highly specialized endeavor.

We present here an alternative proteomic approach that utilizes proteoform intact mass and lysine count determinations, not tandem MS, to reveal proteoforms and proteoform families. A proteoform family, a concept we introduce here, is a set of proteoforms derived from a single gene. Individual proteoforms in a proteoform family frequently differ from one another by single post-translational modifications or amino acid differences but can also differ by larger changes due to splice variation or protein truncation. For example, all of the many different post-translational variants of histone H4 are members of a single proteoform family. A complex proteomic sample such as a cell lysate may contain thousands of proteoform families. We devised a computational process for the determination of the proteoform families present in a complex sample. The families are constructed from knowledge of just two pieces of information for each proteoform, the accurate proteoform mass and the number of lysine residues it contains. Proteoforms are considered to be related, and thus members of the same family, if their lysine counts are identical and their intact masses differ by the mass of known modifications or amino acid changes (in this initial study, we have not yet attempted to include larger changes such as splice variation or protein truncation). Identification of any given member of the family then identifies all members of the family. This initial identification is obtained by matching the accurate mass and lysine count of the experimentally observed proteoform to values calculated from a protein reference database and looking for exact matches within a small mass tolerance. This strategy of using the identification of one proteoform to leverage the identification of many related proteoforms distinguishes this approach from both top-down and bottom-up proteomics, which are based solely upon the identification of individual proteins. In addition, because all members of a family are identified and visualized together, the relative abundances of the related forms are easily compared (see below). The representation and visualization of proteoform families described here nicely parallels related work in the field[19] that connects individual proteoforms and proteoform interaction networks to PTM and disease metadata. This provides an incredible bridge between the experimental process of proteoform identification and the relationship between proteoform observations and the presence of particular disease states.

■ EXPERIMENTAL PROCEDURES

The experimental workflow for identifying proteoforms is straightforward (see the Materials and Methods section in the Supplemental Text). Briefly, accurate proteoform masses are determined by standard LC−MS analysis of undigested protein mixtures in an orbitrap mass spectrometer, and the lysine count is determined using the NeuCode stable isotope labeling by amino acids in cell culture (SILAC) isotopic tagging method.[20] We note that the use of intact mass determination and amino acid count for protein identification has been previously reported.[21−23] We cultured yeast with media containing either of two isotopically heavy forms of lysine: $^{13}C_6{}^{15}N_2$-lysine (+8.0142 Da) or $^{2}H_8$-lysine (+8.0502 Da). These two isotopologues of lysine differ in mass by 36 mDa. Pairs of identical proteoforms produced upon mixing and lysing cells

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

from both cultures have a monoisotopic mass difference equal to 36 mDa times the number of lysines in the proteoform. For these experiments, cells grown in media enriched with "NeuCode Light" and "NeuCode Heavy" lysine are combined in a 2:1 ratio. Experiments here were limited to analysis of proteins below 30 000 Da because of the mass range limitation of the mass spectrometer employed (Thermo LTQ Orbitrap Velos). Cells are lysed, and a soluble protein cleared lysate is prepared, followed by gel electrophoretic separation into 12 molecular weight fractions, which are analyzed by LC-MS.[24] The resultant mass spectra (28 847 in the present study) are processed in a multistep data-analysis pipeline to provide proteoform identifications (an example is shown in Figure 1). The first step in the pipeline is charge-state deconvolution and deisotoping to yield proteoform monoisotopic intact mass values. Protein mass spectra produced by electrospray ionization are highly complex. Each individual protein is observed in multiple different charge states. In addition, the natural abundance C, H, N, O, and S atoms in each proteoform yield multiple different isotopologues. Therefore, mass spectra must be deconvoluted to eliminate the charge-state differences and deisotoped to eliminate the isotopologue effect to obtain a single monoisotopic mass for each proteoform. Next, we paired together mass values that were NeuCode-Light and NeuCode-Heavy isotopologues of one another. The stringent pairing criteria include: a small mass difference of <6 Da, an intensity ratio between 1.4:1 and 6:1 (based on the expected mixing ratio of 2:1), and also observation in the same spectrum and the same charge states (see the Materials and Methods section in the Supplemental Text). This pairing serves two purposes. First, it greatly increases the confidence that the mass values correspond to actual proteoforms from the sample. Second, the number of lysines present in each protein is determined from the mass difference between the doublet peaks for the two proteoform isotopologues using the 36 mDa per lysine conversion factor. Overall, this yielded a set of 70 564 intact masses with associated lysine counts (Supplemental Table S-1), of which 8637 were nonredundant and thus likely to correspond to unique proteoforms (Supplemental Table S-2).

RESULTS AND DISCUSSION

We sought to identify the known yeast proteins to which these 8637 proteoforms correspond. This is not possible to achieve by direct comparison of the UniProt database entries with the experimental data because of the wide variety of possible post-translational modifications, which change the intact proteoform masses. We devised a three-stage strategy to address this problem (Figure 2). In stage 1, experimentally observed proteoforms are identified by pairing them with their theoretical counterparts (experimental−theoretical (ET) pairs); in stage 2, pairs of proteoforms that differ from one another by the mass of well-known protein modifications are identified by pairing them with one another (experimental−experimental (EE) pairs); and in stage 3, all ET and EE pairs sharing a common proteoform are joined together to form proteoform families.

In stage 1 of the strategy, ET pairs are identified by comparing experimental masses with theoretical masses from the UniProt entries having the same lysine count. For each of the 8637 observed proteoforms, we determined the UniProt entries (including single annotated PTMs when present) falling within 500 Da and calculated the differences between the experimental and theoretical proteoform masses. Figure 3A shows a histogram of the results out to 200 Da. The most intense peaks in the histogram correspond to the mass differences associated with frequent protein modifications. Note that several of the major peaks have satellites within one

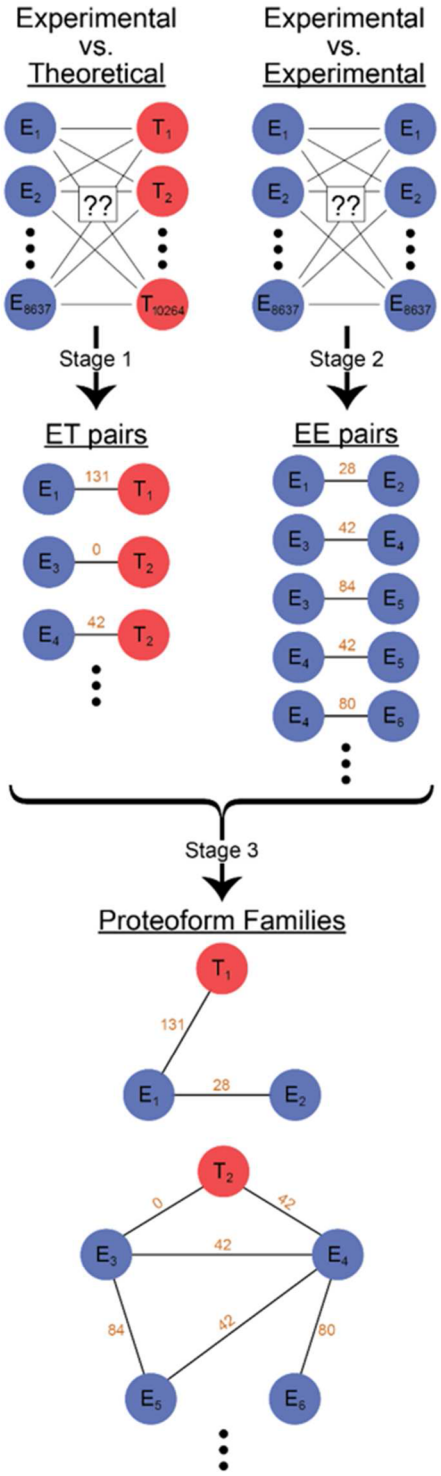


Figure 2. Three-stage strategy for elucidating proteoform families. In the first stage, experimental intact masses, E_n, are compared to theoretical masses, T_n, (having the same lysine count) to create ET pairs for certain mass differences (e.g., 42 Da). In the second stage, EE pairs are similarly generated. In the third stage, the pairs are clustered together to produce proteoform families, two examples of which are shown here.

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

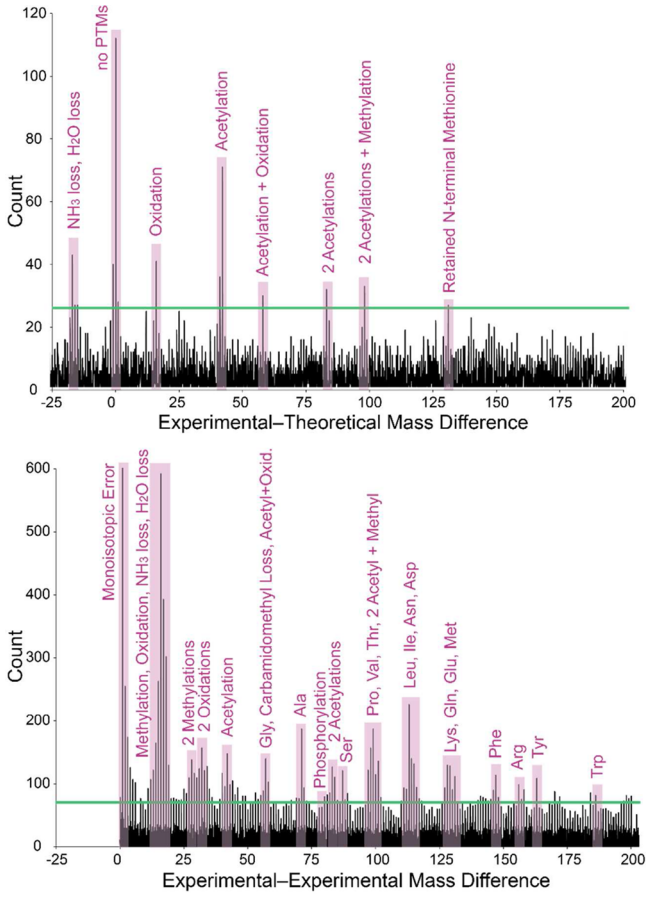


Figure 3. Histograms of observed mass differences. (A) Mass differences between experimental masses and theoretical ones calculated from UniProt entries, which have the same number of lysines. (B) Mass differences between pairs of experimental observations, again stipulating the same lysine count. The most frequently observed mass differences correspond to common PTM or amino acid masses. A total of 31 of the 88 mass differences (highlighted in pink) were directly attributable to known modifications (e.g., oxidation, methylation, and acetylation) or amino acid losses at one of the proteoform termini. Another 34 peaks were adjacent to these 31 primary mass shifts, up to 2 Da away, and were attributed to misassignment of the proteoform monoisotopic mass. Several other mass shifts (e.g., 46 and 72 Da) were included in the construction of proteoform families because they exceeded the threshold but they remain unidentified. These two mass shifts in particular are absent in the compendium of modifications at unimod.org. They could arise from a combination of modifications. An FDR threshold is shown (green line). See also the Supplemental Tables S-9 and S-11 for the complete list of ET and EE mass differences to 500 Da.

or two Da, which are likely due to well-known challenges in the deisotoping of mass spectra of intact proteins.[25] We selected 13 of these mass differences that met an average false discovery rate (FDR) of 21%, ranging from 8 to 35% (see below for a discussion of FDR). This threshold was selected because it captured the major peaks in the histogram corresponding to known prevalent modifications. There were 550 ET pairs identified by this process.

In stage 2 of the strategy, EE pairs are identified by comparing all experimental masses of the same lysine count with one another. For each of the 8637 observed proteoforms, we identified the sets of observed proteoforms having the same lysine count and then calculated all pairwise mass differences within each set. A histogram of the aggregated results for all mass differences below 200 Da is shown in Figure 3B. Peaks highlighted in the histogram include PTMs, amino acid losses, and other protein modifications commonly observed in protein mass spectrometry. We selected the 88 mass differences that met an average false discovery rate (FDR) of 22% (ranging from 5 to 36%; see below for a discussion of FDR). The larger number of significant mass differences observed for EE pairs (88) than for ET pairs (13) is due in part to the multiplicative effect of the monoisotopic errors. For example, we may see a proteoform with a monoisotopic mass of 10 000 Da and a missed monoisotopic mass for that same proteoform at 10 001 Da. The oxidized version of these two forms would have monoisotopic masses of 10 016 and 10 017 Da, respectively. The EE mass differences for all four species would be 1, 15, 16, and 17 Da, with relative intensities of 1:1:2:1. Thus, two actual proteoforms produce four separate peaks in the EE histogram. Stage 2 yielded 11 213 EE pairs.

In the third stage of analysis, proteoform families are formed by joining together all ET and EE pairs sharing a common proteoform. Each pair consists of two nodes (masses of the two proteoforms) and one edge (the mass difference between the two proteoforms). All pairs having a common node are joined together to form discrete proteoform families. This process yielded 1178 proteoform families ranging in size from 2 to 150 members, as displayed in Figure 4A. The proteoform families are represented as collections of nodes and edges, where each node corresponds to a particular proteoform with an associated intact mass and lysine count, and the edges correspond to the mass differences between related proteoforms. The red nodes represent the mass and lysine count of an unmodified (base) protein from a protein reference database (UniProt), the green nodes represent the mass and lysine count of a UniProt-curated post-translational modification of the base reference protein entry (base + PTM), and the blue nodes represent experimental mass and lysine count observations from the yeast lysate sample. The area of each blue node is proportional in size to the number of times that proteoform was observed experimentally, providing a crude measure of abundance. In the simple proteoform family shown in Figure 4B for Negative cofactor 2 complex subunit β, for example, there are four nodes and three edges. The red and green nodes represent the UniProt entries for the base and phosphorylated protein, respectively, and the two blue nodes correspond to the experimentally observed mass and lysine count pairs for both proteoforms. There are two zero Da mass difference edges shown, connecting the UniProt entries with the experimental observations for those proteoforms, and one 80 Da mass difference edge connecting the two experimentally observed proteoforms, corresponding to the mass added upon phosphorylation. Figure 4C–E shows three other proteoform families of increasing complexity, showing multiple methylations of 60S ribosomal protein L12-A, multiple acetylations of Histone H2B.1, and a pattern of amino acid losses from the N-terminal degradation of 60S ribosomal protein L40. The ability shown here to identify and visualize the members of proteoform families provides a powerful and unprecedented new view of proteome complexity at the intact proteoform level, information that is critical to understanding biological systems and pathways.

Figure 5 summarizes the proteoforms and proteoform families identified. Of the total 8637 proteoforms observed, 2378 were not associated with any other proteoform or a

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213–1221

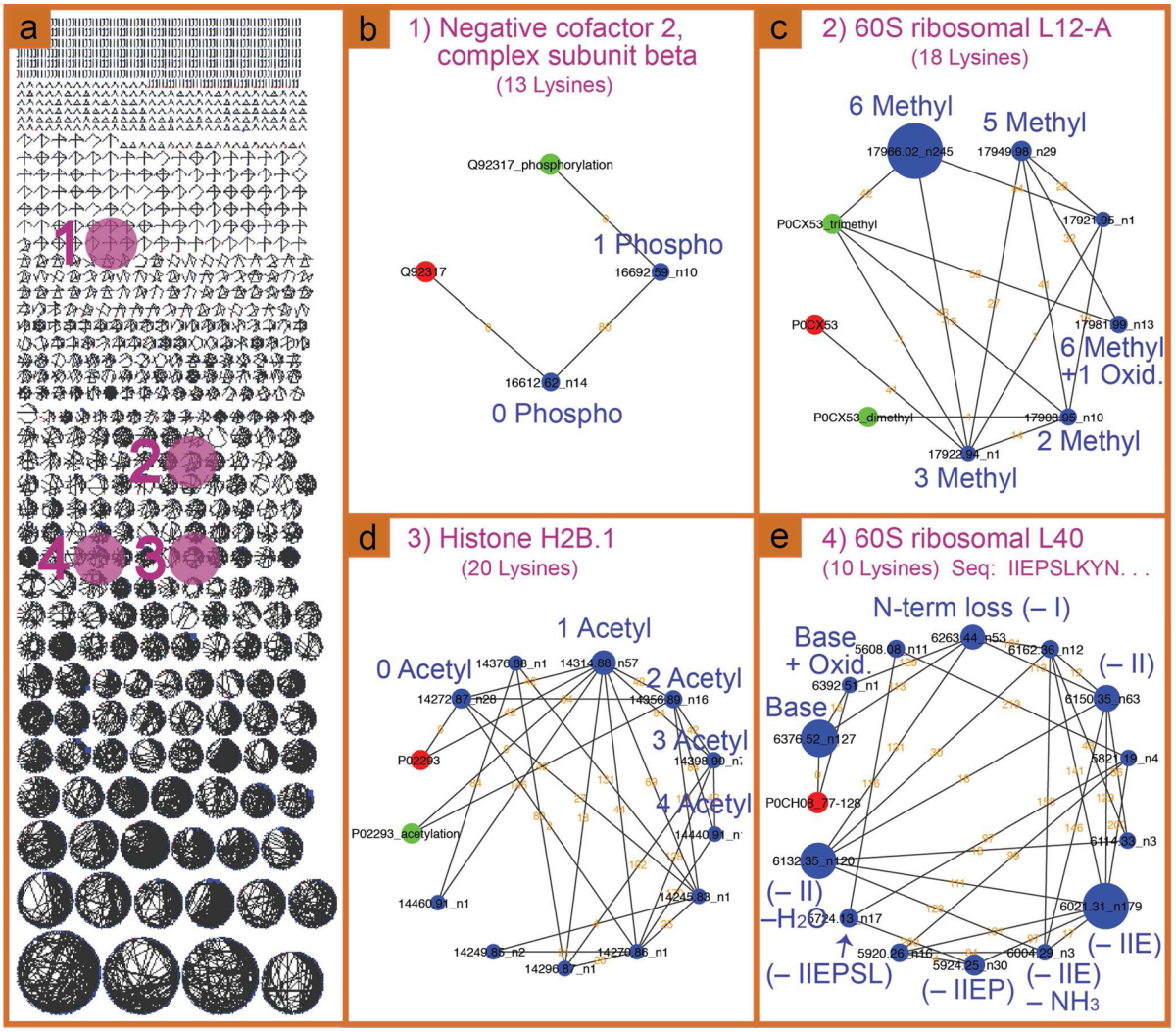


Figure 4. Proteoform families. (A) Display of 1178 proteoform families discovered in this work. (B–E) Expanded views of four example proteoform families. Theoretical unmodified proteins (red nodes) are labeled with their UniProt accession number. Theoretical modified proteins (green nodes) are labeled with their accession number and a PTM known to occur on that protein. Experimentally observed proteoforms (blue nodes) are labeled with their intact mass and the number of times it was detected. The area of each blue node is proportional in size to the number of times that proteoform was observed experimentally; however, to facilitate visualization, all nodes corresponding to 1–10 observations were given the same (minimum) size. Proteoforms are connected by select mass differences (edges) indicated by black lines with orange mass-difference values.

UniProt accession number and hence are not members of families (orphans). The rest of the proteoforms formed 1178 proteoform families composed of 1460 proteoforms belonging to 199 families that correspond to a known protein (i.e., are associated with a single UniProt accession number); 802 proteoforms in 27 families that leave some ambiguity in identification in that they were associated with two or more accession numbers; and the remaining 3997 proteoforms in 952 families that remain unidentified. Of the 70 564 total experimental proteoform observations, 92% belong to one of the 1178 proteoform families. 1216 (14%) of the 8637 proteoforms observed, and 253 (11%) of the 2262 that were also identified, had masses below 5000 Da and thus might be considered as peptides rather than proteins (see Supplemental Table S-2 for a list of all observed and identified proteoform masses, along with histograms showing their distribution as a function of mass). The size distribution of the families is plotted in Supplemental Figure S-1 and shows a roughly exponential decrease in frequency with increasing size. This plot reveals for the first time the number of different proteoforms for a given base protein, providing a new way of assessing the complexity of the entire yeast proteome.

To assess the statistical confidence associated with the identifications, we estimated the false discovery rate (FDR) for the ET and EE pairs. FDR is an estimate of the fraction of false positive identifications in a group of identifications. The strategies for assessing FDR for each pair type are described briefly below and provided in greater detail in the Supporting Information.

FDRs for the ET pairs were determined using a target-decoy strategy, analogous to the widely employed estimation of FDR in bottom-up proteomics.[26] In bottom-up proteomics, the most common method of creating a decoy database for all proteins in an organism of interest is to reverse all of the amino acid sequences. However, this method of creating a decoy database is not useful here because all of the decoy entries would have the same masses and lysine counts as the true target database. We accordingly developed an alternative strategy for the construction of the decoy database. We first concatenated all yeast protein sequences in random order into a single

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213–1221

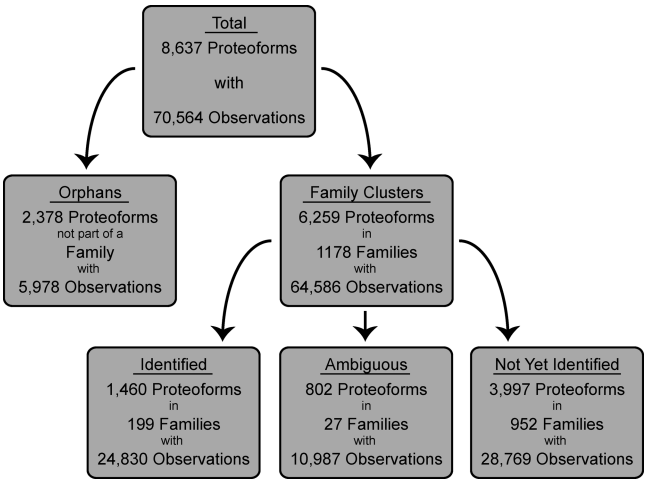


Figure 5. Distribution of observed proteoforms in various types of proteoform families. Most of the observed proteoforms clustered with theoretical or other experimental proteoforms to make families, although some did not (i.e., "orphans"). The proteoform families are categorized as identified, ambiguous, or not yet identified based on containing one, two or more, or zero theoretical accession numbers, respectively. The term "Observation" here refers to each detection of a proteoform intact mass and lysine count in any of the 29 847 mass spectra collected in this study.

continuous string and then divided the string into substrings with lengths equal to each of the known yeast proteins. This yields a decoy database, in which the number and length of decoy protein sequences matches exactly to the known set of yeast proteins, but the masses and lysine counts differ. The database was further expanded to include proteoforms with single post-translational modifications, one for each modification annotated in the UniProt yeast protein database. We created 10 such decoy databases and employed each of them for the stage 1 ET identification to determine the number of experimental−decoy (ED) pairs, which represent false ET connections. The FDR at each mass difference is the ratio of the median number of ED pairs to the number of ET pairs and ranged from 8 to 35%, with an average of 21%. The primary factor driving this high FDR for ET pairs (and for EE pairs below) is mass accuracy, which is limited by the instability and drift in the measurement of intact mass that occurs over the course of the experiment, which requires several days of instrument operation. The false discovery rate is expected to drop with improvements to instrument stability, such as the utilization of a lock-mass standard in the chromatographic buffer for continuous mass calibration.[27]

The target-decoy strategy just described for the estimation of FDR for ET pairs is not applicable to estimation of FDR for EE pairs because no theoretical database is utilized for identifying EE pairs. A different method was needed to estimate the number of false EE pairs at each of the 88 mass differences. We hypothesized that because all true EE pairs are between proteoforms having the same lysine count, we could use mass differences between experimental values having unequal lysine count as a proxy for false-positive connections. To implement this approach, we calculated the mass differences between all experimentally observed proteoforms differing in lysine count by two or more lysines. Because this set of mass differences is vastly larger than the set created when considering only experimental values with the same lysine count, we selected a random subset of size equal to the number of mass differences produced in the EE comparison of Figure 3B. We counted the number of mass differences in this subset in a small window (± 0.04 Da) around each of the selected EE peaks. This count provides an estimate of the number of false EE connections (experimental−false lysine count (EF) pairs) in each peak. The FDR at each mass difference is the ratio of the number of EF pairs to the number of EE pairs and ranged from 5 to 36%, with an average of 22%.

We note that the modest FDR values reported here (21% for ET and 22% for EE) do not compare favorably with either bottom-up proteomics, which commonly reports 1% FDR values for protein identification, or top-down proteomics, which commonly reports 1−5% FDR values for protein identifications. These FDR values for the intact mass and lysine count approach are highly dependent on instrumental factors that can be improved, and therefore, they should not detract from the importance of this new approach to proteoform and proteoform family identification.

We compared the identifications obtained from the intact mass and lysine count strategy with those obtained by top-down proteomics. Briefly, we aggregated yeast top-down search results (Supplemental Table S-3) obtained in our own laboratory (Supplemental Tables S-4 and S-5) with those reported by the Kelleher laboratory in the most comprehensive study published to date.[17] A detailed explanation of this comparison is provided in the Supplemental Text and further supported by additional data found in Supplemental Tables S-6 through S-17. We found 75% agreement between the proteoforms identified by top-down proteomics and the ones identified by the intact mass and lysine count strategy.

It is of interest to note several current limitations of the intact-mass approach to identification of proteoforms and proteoform families, which offer interesting paths forward for the further development of the strategy. The method does not localize PTMs. Localization could possibly be achieved using either bottom-up or top-down mass spectrometry, but neither method guarantees sequence coverage over the region containing the PTM. The strategy will be necessarily more difficult to implement on samples from more complex organisms such as plant and mammalian species because they have larger proteomes and include genetic variation among individuals. Thus, it will be necessary to characterize the sequence variation of the individual under study using large-scale genomic or transcriptomic sequence data to inform and improve the proteomic analysis. Efforts to accomplish this are currently an active area of research, referred to as "proteogenomics".[28] We have used this approach successfully to improve bottom-up proteomic analyses in a variety of mammalian cell lines[29,30] and anticipate that it will be similarly useful for proteoform family analysis. The NeuCode SILAC isotopic tagging strategy employed in this study to provide lysine counts for each proteoform was extremely useful but also limits the approach, as it is not applicable to tissue samples. However, it may be that as comprehensive proteoform databases are established in higher organisms, the lysine-count parameter will become less critical to the identifications and can be replaced by other readily measured or calculated parameters such as chromatographic retention time.[31] Only the rudimentary quantification of proteoforms was accomplished in the current work based on the number of times each mass and lysine count was observed. The accuracy and precision will be greatly improved by using intensity-based measurements or isotopic tagging strategies such as NeuCode SILAC for relative

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

quantification.[20,32] Finally, we believe that there is room for much improvement over the first-generation bioinformatic and biostatistical approaches presented here. For instance, we are devising descriptive statistical approaches that will provide confidence intervals for the likelihood that each individual node (proteoform) is included in the correct family. See the Supporting Information for more in-depth discussion of these current limitations.

We encountered a few interesting phenomena having potential, yet currently unknown, biological significance. First, the process used for determining ET and EE pairs, which involves making a histogram showing the frequency of mass difference values, revealed several frequent but previously unknown differences. These peaks, like those revealed in similar work by ourselves and others,[12] suggest the possibility of unknown protein modifications. Two examples are the peaks at 46 and 72 Da in the EE plot (Figure 3B). We have observed these mass differences in mass-tolerant bottom-up proteomics analyses of yeast. These two particular cases have also been reported elsewhere.[12] We are currently working to interpret them. Second, we observe a considerable number of proteoforms that are missing one or more amino acids from either the N- or C-terminus or both. Proteoforms displaying this behavior were also identified by us and by Kelleher's group[17] using top-down proteomics.

This new strategy of identifying proteoforms from intact mass, lysine count, and clustering into proteoform families serves to complement rather than replace top-down and bottom-up proteomic approaches. We found 1460 proteoforms associated with 199 single accession numbers and an additional 802 proteoforms associated with two or more accession numbers. These numbers compare reasonably well with the most extensive top-down study in yeast to date, which reported 1103 proteoforms associated with 530 accession numbers at 5% FDR, from the same type of sample and gel fractionation.[17] We also compared our results with bottom-up analyses of the same samples, which yielded 2651 protein identifications. We found that the frequency of intact proteoform identifications correlated strongly with the bottom-up protein abundance as determined by spectral counting (Supplemental Figure S-2), indicating that the more abundant proteins are more readily detected in both strategies. Although it is clear that bottom-up analyses are able to identify far more proteins than either intact mass or top-down analyses, they are not able to reveal proteoforms. The intact mass and lysine count strategy could potentially identify more proteoforms than top-down proteomics within a given amount of instrument time due to the intrinsically simpler nature of the data. The intact mass approach is capable of identifying several proteoforms from each high-resolution full spectrum scan, and there are no fragmentation spectra to acquire. However, on the one hand, in top-down mass spectrometry, each identification comes from a high-resolution fragmentation spectrum obtained for a single selected and isolated precursor (intact proteoform). On the other hand, top-down analysis can yield invaluable data that cannot be obtained from intact mass measurements, namely the positional localization of modifications or sequence variations. Furthermore, the proteoform family concept introduced here is not exclusive to intact mass analyses but could easily be applied to top-down proteomics data to identify additional proteoforms. It is thus apparent that the three proteomic approaches are complementary to one another rather than competitive because each is characterized by differing strengths and weaknesses.

Another interesting way of comparing top-down and intact mass approaches is to consider "discovery" versus "scoring" strategies for proteomics. During the human genome project, the initial phase of single nucleotide polymorphism (SNP) analysis was a discovery phase: as the DNA sequence was generated from different individuals, sequence differences were discovered and catalogued, leading over time to vast databases containing millions of genetic variations. Once these variations were known, the need for additional discovery was diminished, and instead, platforms were developed to query samples for already known SNPs[33] or "scoring". We envision a similar transition developing for proteoform analysis, with a "discovery" phase during which proteoforms are identified and catalogued, populating databases that then enable simpler, less expensive, and higher-throughput proteoform "scoring" approaches to be utilized for most biological studies. An early effort at establishing such proteoform databases has recently been initiated by the Consortium for Top-Down Proteomics.[34,35] We posit that the intact-mass approach will function particularly well for scoring proteoforms, and the proteoform family concept will greatly benefit both proteoform discovery and scoring.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b01090.

Figure S-1: Distribution of proteoform family sizes. Figure S-2: Correlation between proteoform identifications and bottom-up proteomic abundance. Table S-1: Complete set of 70 564 experimental observations of the 8637 proteoforms. Table S-2: Complete set of 8637 proteoforms observed by intact mass analysis. Table S-3: Aggregated top-down proteoform identifications from Smith and Kelleher laboratories. Table S-4: Top-down ProSight biomarker search results. Table S-5: Top-down ProSight absolute mass search results. Table S-9: Complete theoretical database. Table S-10: Count of each ET mass difference with accompanying histogram on the next worksheet. Table S-11: Experimental−theoretical (ET) pairs with any of the 13 selected mass differences. Table S-12: Count of each EE mass difference with accompanying histogram on the next worksheet. Table S-13: Experimental−experimental (EE) pairs with any of the 88 selected mass differences. Table S-14: List of observed proteoforms assembled into 1178 proteoform families. Table S-15: FDRs for each of the 13 selected ET mass differences. Table S-16: FDRs for each of the 88 selected EE. Table S-17: Comparison of accession number assignments for proteoforms identified by both intact mass and top-down analyses. (XLSX)

Table S-6: UniProt accession numbers used for bottom-up search. Table S-7: Bottom-up search results (peptide spectral matches). Table S-8: Bottom-up search results (protein identifications). (ZIP)

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

AUTHOR INFORMATION

Corresponding Author

*Phone: 608-263-2594; fax: 608-265-6780; e-mail: smith@chem.wisc.edu.

Author Contributions

L.M.S. conceived of the use of intact mass and lysine count for proteoform identification. M.R.S. and B.L.F. conceived of the schemes to identify EE and ET pairs and the scheme to construct proteoform family diagrams. R.K. prepared samples for analysis. M.S. performed mass analysis. B.L.F. and A.C. processed the raw MS files to generate a list of monoisotopic masses. M.R.S. wrote software to process lists of monoisotopic masses into proteoform families. M.R.S. conceived of the FDR concepts and wrote related software. M.R.S. and B.L.F. analyzed all of the data and selected appropriate cut-offs of EE and ET pairs. M.R.S. and B.L.F. processed all data for tables and figures. L.M.S., M.R.S., B.L.F, and M.S. wrote the manuscript. R.K. and A.C. edited the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health: R01GM103315 and R01GM114292.

ABBREVIATIONS:

PTM, post-translational modification; FDR, false discovery rate; LC−MS, liquid chromatography−mass spectrometry; PSM, peptide spectral match

REFERENCES

(1) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113*, 2343−94.

(2) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73*, 2092−123.

(3) Claassen, M. Inference and validation of protein identifications. *Mol. Cell. Proteomics* **2012**, *11*, 1097−104.

(4) Li, Y. F.; Radivojac, P. Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics* **2012**, *13* (16), S4.

(5) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol. Cell. Proteomics* **2014**, *13*, 339−47.

(6) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509*, 575−81.

(7) Geiger, T.; Velic, A.; Macek, B.; Lundberg, E.; Kampf, C.; Nagaraj, N.; Uhlen, M.; Cox, J.; Mann, M. Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* **2013**, *12*, 1709−22.

(8) Xu, H.; Dephoure, N.; Sun, H.; Zhang, H.; Fan, F.; Liu, J.; Ning, X.; Dai, S.; Liu, B.; Gao, M.; Fu, S.; Gygi, S. P.; Zhou, C. Proteomic Profiling of Paclitaxel Treated Cells Identifies a Novel Mechanism of Drug Resistance Mediated by PDCD4. *J. Proteome Res.* **2015**, *14*, 2480−91.

(9) Smith, L. M.; Kelleher, N. L. Consortium for Top Down Proteomics, Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186−7.

(10) Jenuwein, T.; Allis, C. D. Translating the histone code. *Science* **2001**, *293*, 1074−1080.

(11) Tvardovskiy, A.; Wrzesinski, K.; Sidoli, S.; Fey, S. J.; Rogowska-Wrzesinska, A.; Jensen, O. N. Top-down and middle-down protein analysis reveals that intact and clipped human histones differ in post-translational modification patterns. *Mol. Cell. Proteomics* **2015**, *14*, 3142−3153.

(12) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33*, 743−9.

(13) Omidi, S.; Schreiber, F.; Masoudi-Nejad, A. MODA: an efficient algorithm for network motif discovery in biological networks. *Genes Genet. Syst.* **2009**, *84*, 385−95.

(14) Ye, D.; Fu, Y.; Sun, R. X.; Wang, H. P.; Yuan, Z. F.; Chi, H.; He, S. M. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010**, *26*, i399−406.

(15) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 683−93.

(16) Savaryn, J. P.; Skinner, O. S.; Fornelli, L.; Fellers, R. T.; Compton, P. D.; Terhune, S. S.; Abecassis, M. M.; Kelleher, N. L. Targeted analysis of recombinant NF kappa B (RelA/p65) by denaturing and native top down mass spectrometry. *J. Proteomics* **2015**, *134*, 76.

(17) Kellie, J. F.; Catherman, A. D.; Durbin, K. R.; Tran, J. C.; Tipton, J. D.; Norris, J. L.; Witkowski, C. E., 2nd; Thomas, P. M.; Kelleher, N. L. Robust analysis of the yeast proteome under 50 kDa by molecular-mass-based fractionation and top-down mass spectrometry. *Anal. Chem.* **2012**, *84*, 209−15.

(18) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics* **2013**, *12*, 3465−73.

(19) Ross, K.; Tudor, C. O.; Li, G.; Ding, R.; ÇElen, I.; Cowart, J.; Arighi, C. N.; Natale, D. A.; Wu, C. H. In *Knowledge Representation of Protein PTMs and Complexes in the Protein Ontology: Application to Multi-Faceted Disease Analysis*. Proceedings of the International Conference on Biomedical Ontology (ICBO), Houston, TX, October 6−9, 2014; pp 43−46.

(20) Rhoads, T. W.; Rose, C. M.; Bailey, D. J.; Riley, N. M.; Molden, R. C.; Nestler, A. J.; Merrill, A. E.; Smith, L. M.; Hebert, A. S.; Westphall, M. S.; Pagliarini, D. J.; Garcia, B. A.; Coon, J. J. Neutron-encoded mass signatures for quantitative top-down proteomics. *Anal. Chem.* **2014**, *86*, 2314−9.

(21) Chen, X.; Smith, L. M.; Bradbury, E. M. Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal. Chem.* **2000**, *72*, 1134−43.

(22) Martinovic, S.; Veenstra, T. D.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. Selective incorporation of isotopically labeled amino acids for identification of intact proteins on a proteome-wide level. *J. Mass Spectrom.* **2002**, *37*, 99−107.

(23) Veenstra, T. D.; Martinovic, S.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. Proteome analysis using selective incorporation of isotopically labeled amino acids. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 78−82.

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221

(24) Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L. A robust two-dimensional separation for top-down tandem mass spectrometry of the low-mass proteome. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2183−91.
(25) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* **2010**, *9*, 2772−82.
(26) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55−71.
(27) Staes, A.; Vandenbussche, J.; Demol, H.; Goethals, M.; Yilmaz, S.; Hulstaert, N.; Degroeve, S.; Kelchtermans, P.; Martens, L.; Gevaert, K. Asn3, a reliable, robust, and universal lock mass for improved accuracy in LC-MS and LC-MS/MS. *Anal. Chem.* **2013**, *85*, 11054−60.
(28) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114−25.
(29) 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*, 228−40.
(30) 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*, 2341−53.
(31) Klammer, A. A.; Yi, X.; MacCoss, M. J.; Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **2007**, *79*, 6111−8.
(32) Hung, C. W.; Tholey, A. Tandem mass tag protein labeling for top-down identification and quantification. *Anal. Chem.* **2012**, *84*, 161−70.
(33) Perkel, J. SNP genotyping: six technologies that keyed a revolution. *Nat. Methods* **2008**, *5*, 447−453.
(34) Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolic, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; Leduc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H.; Brodbelt, J. S.; Agar, J. N.; Pasa-Tolic, L.; Kelleher, N. L.; Young, N. L. The first pilot project of the consortium for top-down proteomics: a status report. *Proteomics* **2014**, *14*, 1130−40.
(35) Perkel, J. M. Tearing the top off "Top-Down" Proteomics. *BioTechniques* **2012**, *53*, 75−8.

DOI: 10.1021/acs.jproteome.5b01090
J. Proteome Res. 2016, 15, 1213−1221