analytical chemistry

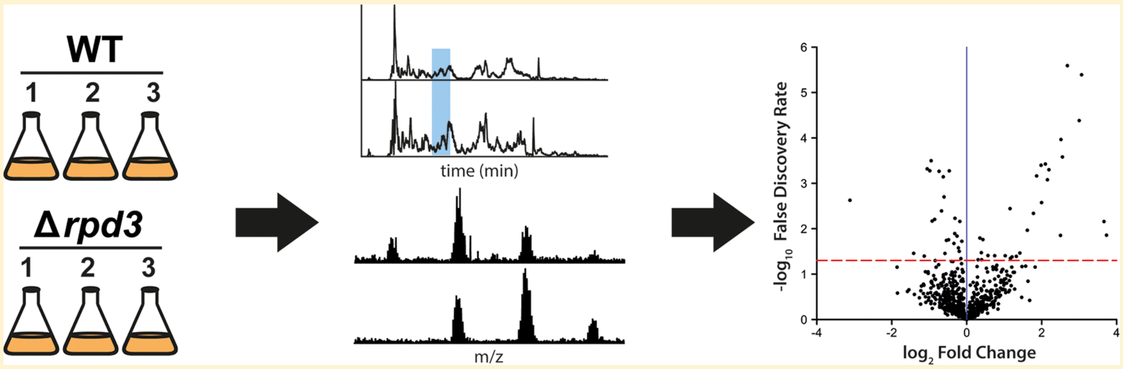
Applying Label-Free Quantitation to Top Down Proteomics

Ioanna Ntai,[†,§] Kyunggon Kim,[†,§] Ryan T. Fellers,[†] Owen S. Skinner,[†] Archer D. Smith, IV,[†] Bryan P. Early,[†] John P. Savaryn,[†] Richard D. LeDuc,[‡] Paul M. Thomas,[†] and Neil L. Kelleher[†,*]

[†]Departments of Chemistry, Molecular Biosciences and the Proteomics Center of Excellence, 2145 N. Sheridan Road, Evanston, Illinois 60208, United States

[‡]National Center for Genome Analysis Support, Indiana University, 2709 E. 10th Street, Bloomington, Indiana 47408, United States

S Supporting Information



ABSTRACT: With the prospect of resolving whole protein molecules into their myriad proteoforms on a proteomic scale, the question of their quantitative analysis in discovery mode comes to the fore. Here, we demonstrate a robust pipeline for the identification and stringent scoring of abundance changes of whole protein forms <30 kDa in a complex system. The input is ∼100−400 μg of total protein for each biological replicate, and the outputs are graphical displays depicting statistical confidence metrics for each proteoform (*i.e.*, a volcano plot and representations of the technical and biological variation). A key part of the pipeline is the hierarchical linear model that is tailored to the original design of the study. Here, we apply this new pipeline to measure the proteoform-level effects of deleting a histone deacetylase (*rpd3*) in *S. cerevisiae*. Over 100 proteoform changes were detected above a 5% false positive threshold in WT vs the Δ*rpd3* mutant, including the validating observation of hyperacetylation of histone H4 and both H2B isoforms. Ultimately, this approach to label-free top down proteomics in discovery mode is a critical technical advance for testing the hypothesis that whole proteoforms can link more tightly to complex phenotypes in cell and disease biology than do peptides created in shotgun proteomics.

Since the development of soft ionization techniques over 20 years ago, mass spectrometry (MS) has become the method of choice for untargeted protein analysis. However, owing to the difficulty of *intact* protein analysis by MS (arising from both hardware and software challenges[1]), the vast majority of proteomics research has been developed and conducted using a "bottom-up" approach, where proteins are first digested into constituent peptides prior to MS analysis.[2] Top down proteomics describes the process for identification and characterization of intact protein forms (*i.e.*, proteoforms[3]) by mass spectrometry without the preanalytical variables introduced by the digestion step itself.[4−7] While the field of quantitative bottom-up proteomics has undergone multiple advances in both labeled and label-free quantitation,[8−10] similar advances in the field of top down proteomics to analyze hundreds or thousands of proteoforms in quantitative fashion are not available at present.

A major milestone in top down proteomics has been the publication of several studies showing the high-throughput identification of thousands of distinct proteoforms within bacterial and mammalian cell lysates using modern high resolution MS instrumentation coupled to nanocapillary liquid chromatography (nLC).[11−14] In studies from our group, a top down proteomics pipeline was established to reduce sample complexity using an off-line size-based separation (gel elution liquid fractionation entrapment electrophoresis or "GELFrEE") followed by capillary LC-MS/MS of intact proteins to be detected in high resolution Orbitrap (or FT-ICR) MS instrumentation. While these studies afforded a rich look into the mammalian cell proteome at the intact proteoform level, a similar pipeline for quantitative label free top down proteomics has not yet been developed.

Several laboratories have established approaches for the targeted quantitation of whole proteins within a mixture of limited complexity. One straightforward approach uses the

Received: January 27, 2014
Accepted: April 28, 2014
Published: May 7, 2014

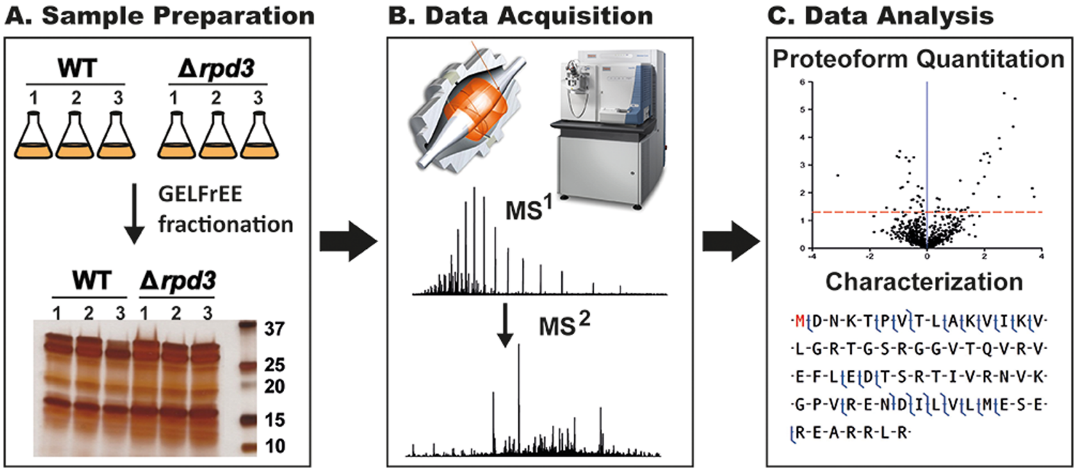


Figure 1. Overall workflow for label free quantitation of whole proteoforms using top down proteomics. With prescribed numbers of biological and technical replicates, a size-based fractionation of whole proteins from a complex proteome is performed (A), followed by randomized LC-MS/MS runs (B) and integrated application of a standard linear model for statistical evaluation of results such as the volcano plot (panel C, top) where each dot represents a proteoform that can be identified and characterized (panel C, bottom).

measurement of intensity ratios for multiple, coeluting proteoforms to establish relative quantitation within a single sample.[15] Since this intraspectrum quantitation holds all of the information necessary within one instrument data file, it is somewhat immune to the variability inherent within large multisample quantitative studies. Examples of this technique include the work of Dong et al. on the quantitation of cardiac troponin I proteoforms in heart tissue in patients with congestive heart failure[16] and the work of Chamot-Rooke et al. in the quantitation of *N. meningitides* type IV pili proteoforms.[17]

To perform proteome-wide quantitation, several groups have taken both *in vivo* and *in vitro* labeling approaches with varying success.[18−21] While both approaches are well established for comparative proteomics as they minimize technical variation by mixing samples prior to analysis, there are a number of challenges hindering their development and implementation on a wide-scale basis.[9,10] Du et al. used an *in vitro* differential cysteine labeling strategy to quantify intact proteins from yeast grown under aerobic and anaerobic conditions. Although, in theory, this strategy allows MS1-based quantitation of protein pairs, they found that the differential tags altered chromatographic retention time, thus interfering with intraspectrum quantitation.[19] More recently, Hung et al. used *in vitro* tandem mass tag (TMT) labeling with isobaric tags to perform MS2-based intrascan quantitation.[21] An advantage of isobaric tags is that labeled protein pairs should have identical chromatographic profiles. However, because this approach uses MS2 fragmentation data for quantitation, its accurate implementation requires only one precursor ion be selected for fragmentation, which is often not the case in top down proteomics of complex samples.[14] *In vivo* labeling with stable isotopes shares the chromatography advantage of *in vitro* isobaric labeling but circumvents the requirement for single precursor ion isolation. In a past study, our group implemented $^{14}N/^{15}N$ labeling and quantified over 200 protein pairs from yeast grown in the presence or absence of oxygen at the intact protein level.[18] More recently, Collier et al. applied this strategy to human embryonic stem cells grown in culture.[20] In all cases, implementation of this strategy required the ability to label cells *in vivo*, thus limiting the technique to applications where feedstocks can be manipulated, such as cell and tissue culture. As one of the most pressing goals of comparative top down proteomics is the discovery of biomarkers in clinical research (which precludes metabolic labeling of proteins),[22] it is necessary to develop a statistically valid label-free approach.

Several groups have applied label-free quantitation to comparative top down experiments on a few, targeted proteoforms. Yates' group has pioneered "differential mass spectrometry" (dMS) to perform relative quantitation of proteoforms of apolipoprotein C−III within high-density lipoprotein particles.[23−25] In another example, Taylor et al. utilized charge-state abundances from MS1 spectra and DeCyder software to calculate the relative abundances of large, secreted peptides (up to 60 amino acids) from cell culture in stimulated and unstimulated cells.[26] In both cases, analysis was limited to a few proteoforms and statistical assessment was performed using a traditional Student's *t* test. While acceptable for comparing two biological conditions across a set of technical replicates, Student's *t* test is insufficient to address the many sources of technical variation inherent in complex, multilevel comparative proteomic studies; ANOVA is required to correctly handle multiple levels of variation for quantitative proteomics run in discovery mode.

S. cerevisiae is often the model system to benchmark new proteomics technology; it is readily grown to large quantities and well characterized at the protein level.[27−29] Additionally, there are a number of knockout strains available enabling the global proteomic profiling resulting from the loss of a single gene.[30] One such genetic mutant is the *rpd3*Δ::*KANMX* strain. The *rpd3* gene encodes a histone deacetylase; its deletion has been shown to increase the acetylation levels of all core histones.[31] Additionally, *rpd3* deletion has been shown to increase yeast doubling times by nearly 2-fold[32] and have other global effects owing to a lack of epigenetic regulation.[33]

Here, we have expanded a top down proteomics platform[12,14] to include label-free quantitation of proteoforms <30 kDa for discovery mode research (Figure 1). We developed the platform using a hierarchical linear statistical model capable of handling multiple levels of variation inherent to comparative proteomics experiments. First, we present proof of principle for this analysis through the standard addition of protein standards

Table 1. Experimental Description and Coefficients of Variation for Spiking Standards into a Yeast Lysate at Three Defined Levels (See Figure S1 for Additional Results)

	levels (pmol/10 μL injection)			CV uncorrected			CV normalized		
standard	1X	3X	12X	1X	3X	12X	1X	3X	12X
ubiquitin, bovine	0.14	0.41	1.6	46%	21%	43%	15%	14%	9%
myoglobin, equine	1.1	3.3	13	24%	19%	49%	12%	8%	11%
trypsinogen, bovine	0.48	1.5	5.8	44%	21%	45%	12%	15%	20%
carbonic anhydrase II, bovine	0.64	1.9	7.7	N/Q[a]	N/Q	N/Q	N/Q	N/Q	N/Q

[a] Not quantitated.

to a complex yeast proteome background. We then applied this top down quantitative platform to wild type vs Δ*rpd3 S. cerevisiae* and quantified 120 proteoform differences (54 from the nucleus, 66 from the cytosol) with false discovery rates (FDRs) for the quantitation ranging from 5% to better than 0.0001%. A similar SILAC study was performed by Henriksen et al. in 2012. While the main focus of the paper was differential acetylation profiling by bottom-up proteomics, our results are concordant with those (*vide infra*).[34] To our knowledge, this work reports the first analytical platform for the large-scale label-free quantitation of whole proteins in complex mixtures using top down proteomics.

■ MATERIALS AND METHODS

Yeast Growth and Sample Preparation. Single colonies of wild type *Saccharomyces cerevisiae* S288c BY4742 and the *rpd3*(YNL330C) deletion mutant (*rpd3*Δ::*KANMX*) were picked and were inoculated into 5 mL each of liquid YPD media without and with 0.2 g/L G-418, respectively. After overnight incubation (250 rpm @ 30 °C) and centrifugation at 3000 rpm for 10 min, each pellet was gently resuspended with 1 mL of liquid YPD and was inoculated into 250 mL of YPD and YPD+G-418. Cells were harvested at OD_{600} = 0.7 by centrifugation at 3000 rpm for 20 min. Supernatants were discarded, and each cell pellet was washed with distilled water. The mass of each cell pellet was measured before storage at −80 °C.

Lysis and extraction of *S. cerevisiae* was performed using YPER (ThermoPierce, Rockford, IL) supplemented with 5 nM microcystin, 500 μM 4-(2-aminoethyl)benzenesulfonyl fluoride (AEBSF), 100 mM sodium butyrate, and 100 mM dithiothreitol (DTT) at 2.5 mL/g wet cell weight according to the manufacturer's protocol. After each centrifugation step, the supernatant was saved as the cytosolic fraction, and the protein concentration was determined. To isolate nuclear proteins, each pellet was resuspended using 30 mL of 100 mM of sodium butyrate and centrifuged at 18 000*g* for 10 min at 4 °C to remove the YPER. Next, an acid/urea extraction of the histone fraction was performed by adding 2.5 volumes of 8 M deionized urea with 0.4 N of sulfuric acid and vortexing for 5 min and extraction on ice for 30 min. C4 solid phase extraction (Bakerbond C4, J.T Baker) was performed, and after washing the column with 30 mL of 0.1% trifluoroacetic acid (TFA) in water, the sample was eluted with 3 mL of 0.1% TFA in 60% acetonitrile. Each eluted fraction was dried and reconstituted with 1.0% sodium dodecyl sulfate (SDS) solution to quantify the amount of proteins using bicinchoninic acid (BCA) assay (Pierce, Rockford, IL).

Four hundred micrograms of total protein was prepared for each lane of GELFrEE per manufacturer's instructions (Expedeon, Cambridgeshire, UK, GELFREE 8100). Each biological replicate was separated on a single lane of an 8%T GELFrEE cartridge (6 lanes total). Fraction 1 was collected for LC-MS/MS analysis. Ten microliters of the 150 μL fractions was used for conventional SDS-PAGE analysis and silver stain visualization (Figure 1). SDS was removed by methanol/chloroform/water extraction.[35] Proteins were resuspended in 40 μL of Buffer A (95% H_2O, 5% AcN, 0.2% FA). Samples were centrifuged for 10 min at 21 000*g* at 4 °C prior to injection.

Standard Spike Experiment. Known amounts of a "Top Down Standard" containing ubiquitin (Sigma-Aldrich, U6253), trypsinogen (Sigma-Aldrich, T1143), myoglobin (Sigma-Aldrich, M5696), and carbonic anhydrase (Sigma-Aldrich, C2522) were added to a fixed background of yeast wild-type nuclear lysate (Table 1). Data analyses were performed as described below but adjusted to analyze a simpler experiment. Samples were each injected four times to observe technical variance.

LC-MS/MS Parameters. Resuspended protein fractions (5 μL) were injected onto a trap column (150 μm ID × 2 cm) using an autosampler (Dionex). A nanobore analytical column (75 μm ID × 15 cm) was coupled to the trap in a vented tee setup. The trap and analytical columns were packed in-house with polymeric reverse phase (PLRP-S, Phenomenex) media (5 μm, 1000 Å pore size)[36] and connected to 15 μm nanoelectrospray tips (New Objective, Waltham, MA). A Dionex Ultimate 3000 RSLCnano system was operated at a flow rate of 2.5 μL/min for loading onto the trap. Proteins were separated on the analytical column and eluted into the mass spectrometer using a flow rate of 300 nL/min and the following gradient: 5% B at 0 min, 15% B at 5 min, 55% B at 55 min, 95% B from 58 to 61 min, 5% B from 64 to 80 min. Solvent A consisted of 95% water, 5% acetonitrile, and 0.2% formic acid, and solvent B consisted of 5% water, 95% acetonitrile, and 0.2% formic acid.

Mass spectrometry data were obtained on an Orbitrap Elite mass spectrometer fitted with a custom nanospray ionization source. The MS method included the following events: (1) FT scan, four microscans, *m/z* 500−2,000, and resolution 100 000 and (2) data-dependent MS/MS on the top two peaks in each spectrum from scan event 1 using higher-energy collisional dissociation (HCD) with normalized collision energy of 25, isolation width 50 *m/z*, four microscans, and detection of ions with resolving power of 60 000. Dynamic exclusion was enabled with a repeat count of 2, a repeat duration of 120 s, and an exclusion duration of 5000 s. Automatic gain control (AGC) was set to 1E6 ions, and maximum injection time was set to 1 s for both MS^1 and MS^2. Advanced signal processing was turned on, and data were collected in reduced profile mode. A 15 V offset in the source was used over the entire experiment. The capillary temperature was 320 °C and a spray voltage of 1.8 kV.

Data Processing. All data files in the quantitation portion of the platform were processed using a collection of in-house tools to automate data analysis. Files were analyzed for

Quantitation Mass Targets (QMTs) using a moving spectral average and deisotoped with Thermo Fisher's Xtract algorithm at a signal-to-noise value of 6. All QMTs were then binned by mass (8 ppm) and retention time (8 min) to reduce data redundancy. Intensities were normalized using the average total ion chromatogram intensity for each technical replicate. Finally, the QMTs were grouped and artifactual ±1 Da deisotoping errors were removed. Final QMTs were stored within a SQLite reporting database.

Once a set of QMTs was determined, a quantitative algorithm was applied to determine an appropriate intensity value for each QMT across each data file. First, the isotopic distributions of all theoretical charge states of that QMT were generated. These distributions were then used to match against observed spectral data and return intensity estimates for each scan. Finally, the intensity estimates were aggregated across all scans and charge states to report one intensity value for each data file and QMT. These data were provided as a text file for further statistical processing (*vide infra*). At this point, QMTs represent individual, yet uncharacterized, proteoforms.

For proteoform identification and characterization, our conventional top down proteomics pipeline was used as previously described.[12,14,37] Briefly, m/z data for each precursor/fragmentation scan pair were converted to monoisotopic neutral mass values using ProSightHT within ProSightPC 3.0.[38] Data were used to search an annotated *S. cerevisiae* database (stored as a.pwf file called "server_-yeast_complex_Apr3_2013") which was built against UniProt release 2013_04. Mass tolerances for precursor ions were set to 10 ppm. A 10 ppm mass tolerance was also used for the fragment ions. Low confidence proteoform identifications were excluded by requiring those hits arising from an absolute mass search to have an E-value below 1×10^{-4}.[14] A more stringent E-value cutoff of 6×10^{-5} (corresponding to a P-value cutoff of 9×10^{-8}) was applied to hits derived from a biomarker search as previously reported.[39]

Statistical Analysis. Proteoforms may not be observed in all samples; this creates the "missing values" problem in label-free methods. To address this, intensity data on the occurrence of putative proteoforms (QMTs) were tabulated, and those not occurring in at least 50% of all data files were excluded from further analysis. This removed 67% and 83% of potential QMTs in the nucleus and cytosol, respectively. Intensity values for the remaining QMTs were $\log_2$-transformed so that differences in estimated treatment-level intensities could be interpreted on a fold-change scale. Two separate ANOVA analyses were performed: ANOVA-1 and -2. For the first analysis, ANOVA-1, intensity levels for each QMT were standardized to Z-scores across all samples. ANOVA-2 used unstandardized intensity values. ANOVA-1 was used to test the statistical significance of QMT intensity changes between the wild type and Δ*rpd3* mutant strains, while ANOVA-2 was used to estimate the size of the effect (expressed as fold-change). In both analyses, a hierarchical linear model was employed as the general statistical approach. The fixed effect hierarchical linear model allows for nested effects and can be expressed as $I_{ijk} = \mu + A_i + B_{j(ik)} + C_{k(ij)} + \varepsilon_{ijk}$. In ANOVA-1, "$I$" represents the QMT intensity Z-score, while in ANOVA-2, this represents the $\log_2$-transformed intensity. In both models, μ is the true mean, A is the treatment factor levels (wild type and Δ*rpd3*), B is the biological replicates, C is the technical replicates, and ε is the residual variance. QMTs showing significant treatment*biological replicate effects were excluded from further analysis. In ANOVA-1, all *p-values* were corrected for multiple testing at a false discovery rate of $\alpha = 0.05$.[40] All statistical analyses were performed within SAS 9.4, (SAS Institute, Cary NC).

RESULTS AND DISCUSSION

Quantifying Known Protein Abundances within *S. cerevisiae* Lysates. As a proof of principle, we first performed an experiment in which we spiked a known amount of a set of intact protein standards into a yeast nuclear protein extract background at three known levels. To test the fitness of our approach to perform label-free quantitation, we examined the precision and accuracy of the method using three standard proteins (ubiquitin, myoglobin, and trypsinogen; Table 1). Normalization to the total ion count observed in the LC-MS run proved valuable for improving the precision and accuracy of quantitation in the top down proteomics data set. While uncorrected intensity data showed a coefficient of variation (CV) range of 19−49%, normalization reduced that to a range of 8−20% (Table 1, Figure S1). The accuracy of the method varied. For a 3-fold change (3X vs 1X), we observed a range of 2.2−3.4-fold. For a 4-fold change (12X vs 3X), we observed a range of 2.2−3.3-fold. For a 12-fold change (12X vs 1X), we observed a range of 6.8−11.4-fold. Pairwise Student's t tests for each of these three comparisons were significant at $\alpha = 0.05$.

Applying the Method in Comparative Fashion: WT vs Δ*rpd3 S. cerevisiae*. To evaluate the quantitative platform in an unknown system, we employed a comparative proteome analysis of wild type vs Δ*rpd3 S. cerevisiae* (depicted in Figure 1). A 2 × 3 × 7 study design (*i.e.*, two states, three biological replicates, and seven technical replicates) was established, and the cytosolic and nuclear fractions were assessed separately. Proteins in these compartment extracts were first separately resolved into one simple fraction ranging from 3.5−30 kDa using GELFrEE followed by LC-MS/MS as described above.

Next, we applied the hierarchical linear model to quantify intact proteoforms within our yeast experiment. A random effects model was applied to this experiment, and the results are shown in Figure S2. The majority of the variance was confined to the residual term, "ε", of the hierarchical linear model. The contributions of the technical variation from LC-MS procedures was quite small in comparison (Figure S2). To gain a birds-eye view of proteoform-level changes, we performed an unsupervised clustering of Z-scores at the biological replicate level (Figure 2). This data representation provides a visual depiction of the high reproducibility of the approach and its ability to striate different sets of proteoforms by their response to deletion of the *rpd3* gene.

Using the method detailed above, we created a volcano plot which represented each proteoform (*i.e.*, QMT) as a function of estimated effect size (in $\log_2$ fold-change) and the statistical confidence (the FDR) that there was a difference in normalized intensity between the two states "wild type" and "Δ*rpd3*" (Figure 3). As expected, masses fall to each side of the line indicating "no change," where to the left are found proteoforms more highly expressed in wild-type than in the Δ*rpd3* mutant. Proteoforms identified to the right were upregulated in Δ*rpd3* as compared to WT. Of the 838 QMTs detected in total, 120 of them showed a statistically significant intensity change between wild type and Δ*rpd3* at or below the 5% FDR threshold, with the most confident of these approaching an instantaneous FDR value of 1×10^{-6} (Figure 3). Overall, the nuclear and cytosolic compartments showed changes in 54 and 66 proteoforms, respectively (Figure 3A,B and Tables S1−S4).

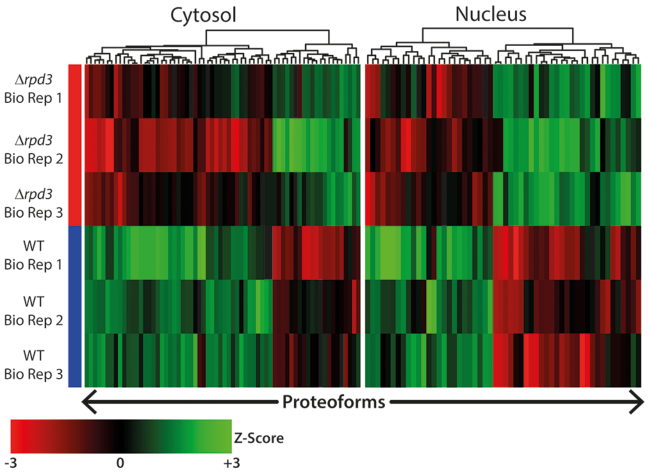


Figure 2. Hierarchical clustering of each of 120 proteoforms found to change significantly (*i.e.*, 5% FDR or better) in the Δ*rpd3* mutant vs WT strains separated by nucleus and cytosol. The clustering diagram used the Z-score approach to score each proteoform across averaged technical replicates, applied to biological replicates and then clustered via an unsupervised approach. Notice that although the clustering was unsupervised, biological replicates from all wild type and Δ*rpd3* runs cluster together.

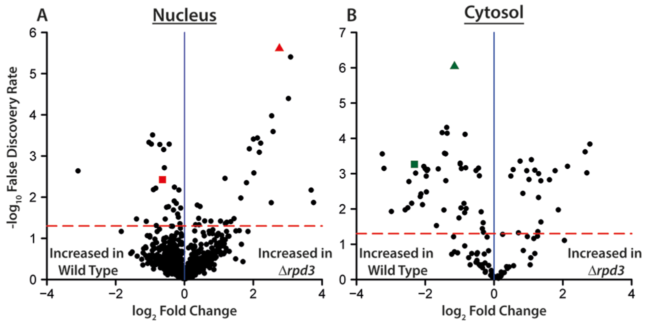


Figure 3. Volcano plots generated to compare the WT vs Δ*rpd3* strains of *S. cerevisiae* S288c for GELFrEE fractions from the (A) nuclear and (B) cytosolic cellular compartments having 54 and 66 proteoforms, respectively, below the 5% FDR threshold (dotted red line). Assignment of analytical variation is shown in Figure S1. In Figure 4, the quantitation of four of these proteoforms is explored in greater detail. The corresponding data points are highlighted above—red ■ = diacetylated histone H4, red ▲ = triacetylated histone H4, green ▲ = N-acetylated ZEO1, green ■ = N-acetylated + phosphorylated ZEO1 (phosphorylation localized between Q12 and T48).

These data demonstrate that the integrated method and statistical model are capable of detecting significant differences in proteoform abundance among treatment groups. Of the QMTs confidently found to change, we used a combination of tandem MS and intact mass tag (IMT)[41] information to map each QMT to a proteoform. Of the 120 QMTs, 71 were unambiguously identified as proteoforms using MS/MS information obtained during the LC-MS runs. An additional five were less-confidently identified using the IMT approach with a 10 ppm tolerance to match protein mass tags obtained from a prior publication reporting >900 yeast proteoforms.[39] With these data, 63% of QMTs were confidently identified as proteoforms while the platform was run in discovery mode.

Within the cytosolic fraction, three of the 66 confidently changing QMTs were related to the protein ZEO1 (UniProt Accession Q08245). ZEO1 is a peripheral membrane protein implicated in the cellular stress response.[42] Our qualitative analysis identified three different proteoforms. Unmodified ZEO1, N-acetylated ZEO1, and N-acetylated + phosphorylated ZEO1 (phosphorylation localized between Q12 and T48). Of the confident QMTs, our quantitation platform detected a significant change between Δ*rpd3* and WT for both the N-acetylated ZEO1 (two QMTs map to one proteoform) and N-acetylated + phosphorylated ZEO1 (Figure 4A, Tables S2 and S4). The abundance of the N-acetylated proteoform was 2.3-fold lower in the Δ*rpd3* mutant as compared to the wild type (most confident instantaneous FDR = 1 × 10^{-6}). The doubly modified proteoform's abundance was reduced by a factor of 5.0-fold in the Δ*rpd3* mutant as compared to the wild type (instantaneous FDR = 5 × 10^{-4}). Most of the proteins seen to be downregulated in mutant cytosol (Table S2) are involved in the stress response and glycolysis.[43,44]

In the nuclear fraction, several of the most confidently changing proteoforms belong to the core histone family. The analysis showed a general hyperacetylation of histone H4 (UniProt Accession P02309) within the Δ*rpd3* mutant as compared to WT (Figure 4B, Tables S1 and S3). This hyperacetylation is evidenced by increased QMT abundance of the triacetylated proteoform in Δ*rpd3* as compared to WT (8.4-fold change, instantaneous FDR = 4 × 10^{-6}). Simultaneously, the diacetylated proteoform was found to be downregulated in Δ*rpd3* as compared to WT (0.6-fold change, instantaneous FDR = 4 × 10^{-3}). ANOVA analyses of the monoacetylated and tetraacetylated proteoforms were not meaningful because intensity values were not present in one of the treatment levels. These "single-state" cases are easily extracted from the intact proteoform area measurements, displayed as boxplots in Figure 4. While the data within this figure show only one technical replicate of each treatment level, similar histone H4 proteoform PTM patterns were seen in all 42 individual technical replicates as a function of treatment level (Figure S3). Given Rpd3's known function as a histone deacetylase and previous reports in both yeast and human cells, the hyperacetylation of histone H4 is not surprising.[31,45] Among other core histones, three proteoforms of histone H2B show significant quantifiable differences for both of its two distinct gene products, H2B.1 and H2B.2 (UniProt Accessions P02293 and P02294). Interestingly, histone H2B.1 shows a more confident degree of hyperacetylation than does histone H2B.2 (Figure 5 and Table S1).

Proteoform-level quantitation offers some advantages over its peptide-level counterparts. Chief among those is the coverage of multiple and diverse modifications afforded by measuring the whole protein. In the cases demonstrated in Figures 4 and 5, peptide-level measurements would have great difficulty to simultaneously quantify correlated changes in multiple modifications in a proteotypic (gene-specific) fashion. For example, histone H2B.1 and H2B.2 are 97% identical in protein sequence and contain a large number of lysine residues. Tryptic digestion would sever the linkage between modifications co-occurring on the same molecule and would provide an unclear view of isoform-specific regulatory events (particularly prevalent in higher eukaryotes). To compare our results to a recent SILAC study looking at the effects of global changes upon deletion of *rpd3*, over 60% of the proteins quantified in this study map to differentially expressed peptides in that report.[34]

Future directions of this platform include its extension to more complex experimental designs and its packaging within a

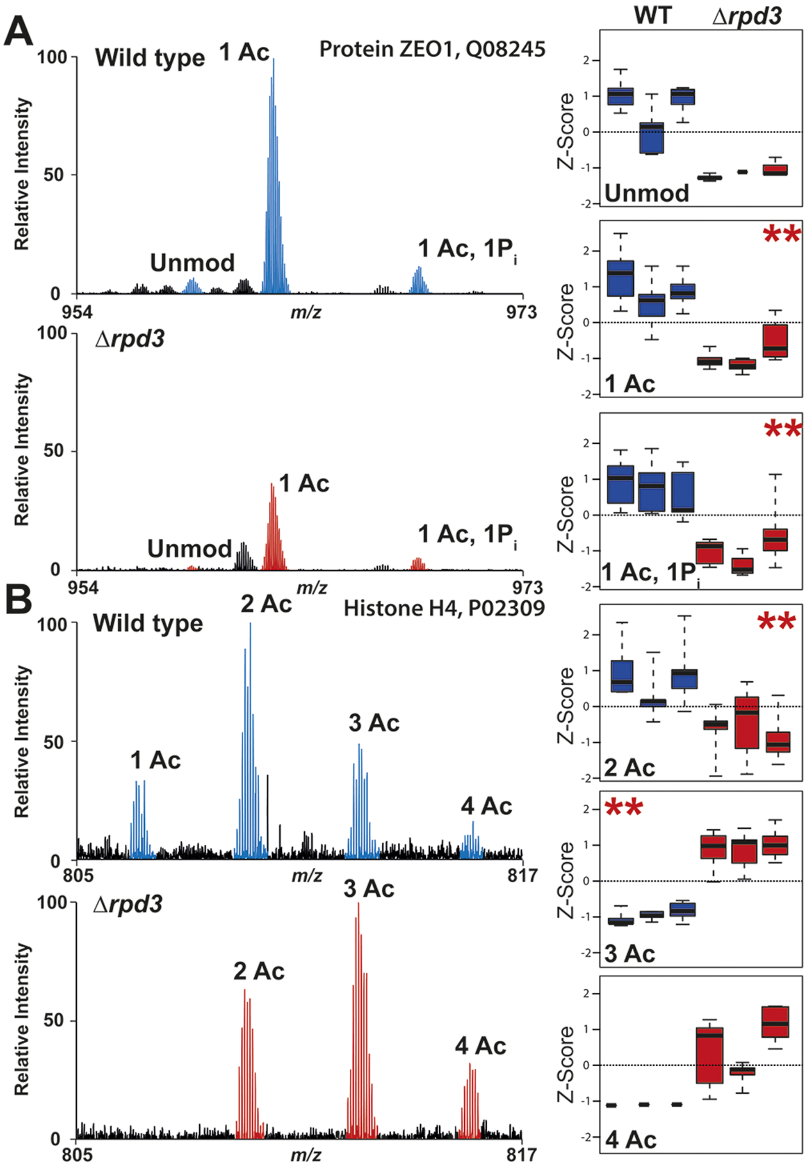


Figure 4. (A) Quantitation of three proteoforms from the cytosolic protein ZEO1. Box and whisker plots are presented at the biological replicate level of Z-scores for each of three proteoforms are shown along the right-hand side. The singly acetylated ZEO1 and the acetylated + phosphorylated ZEO1 show significant changes; the unmodified form was not considered within the analysis because it was observed in fewer than 50% of the technical replicates. (B) Quantitation of histone H4 proteoforms. Box and whisker plots (panels at right) are presented for the histone H4 proteoforms with 2, 3, and 4 acetylations. Again the monoacetylated and tetraacetylated proteoforms were not considered within our analysis because they was observed in <50% of the technical replicates. In all cases, mass spectra are the sums of individual scans across its full elution time within a single technical replicate. The symbol ** indicates significant ($p < 0.05$) proteoform abundance changes as reported by our platform.

software solution. We conclude that our platform is capable of label-free quantitation of intact proteins using top down proteomics and hierarchical linear modeling. We also will extend this method to more complex samples to determine if proteoforms can provide insight into complex phenotypes observed in the human population.

■ CONCLUSION

We forward this approach to demonstrate that quantitative top down proteomics is possible to do in multitarget discovery space. In a model yeast system (with peak finding and statistical approaches implemented with stringency in mind), confidence values up to one in a million were possible to obtain for quantitative measurements of whole proteins in the <30 kDa regime. Extension to primary material from human population studies will likely reduce this level of confidence relative to the clonal yeast population used here. The statistical power generated by the method will improve with better reproducibility of laboratory protocols and fine-tuning of data pipelines. However, the technology is now ready to employ to test the hypothesis that whole proteins can correlate tightly to overall human phenotypes in disease populations. One value proposition is that proteoform discovery and validation will provide robust, protein-based biomarkers that can detect disease early and guide the development of therapeutics in the future of 21st century biomedicine. Such activities have commenced using the technology described and validated in this work.

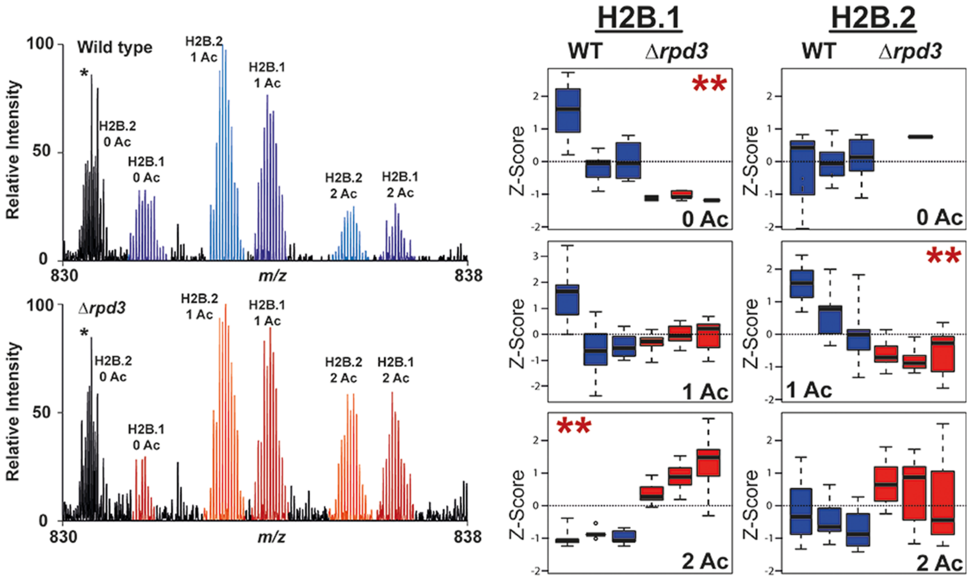


Figure 5. Quantitation of histone H2B.1 and H2B.2 proteoforms across the WT vs Δ*rpd3* strains of *S. cerevisiae* S288c. H2B.1 proteoforms are shown in dark blue and red respectively between WT and Δ*rpd3*. H2B.2 proteoforms are shown in light blue and orange respectively between WT and Δ*rpd3*. Box and whisker plots at right show the significant hyperacetylation of histone H2B.1 (decrease in abundance of unacetylated and increase in abundance of diacetylated) but fail to show a similar trend for histone H2B.2 (although levels of the monoacetylated H2B.2 are decreased within Δ*rpd3*).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures S1−S3 and Tables S1−S4. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +1 847-467-4362. Fax: +1 847-467-3276. E-mail: n-kelleher@northwestern.edu.

Author Contributions

[§]These authors contributed equally. The manuscript was written by P.M.T., N.L.K., I.N., and J.P.S. with contributions from all authors. R.D.L. developed the initial quantitative framework. I.N. and K.K. performed experiments. R.T.F., O.S.S., P.M.T., I.N., K.K., B.P.E., A.D.S., and R.D.L. performed data analysis. All authors have given approval to the final version of the manuscript.

Funding

The following grants are acknowledged: NIH R01 GM067193, NIH P30 DA018310, NSF ABI-1062432, NIH HHSN261200800001E.

Notes

The authors declare the following competing financial interest(s): Qualitative protein analyses were performed in ProSightPC, a product commercialized by the Kelleher Research Group.

■ ACKNOWLEDGMENTS

The authors acknowledge the past work of Dr. Lihua Jiang on yeast histone mass spectrometry. The project described was supported by Award No. DA018310 from the National Institute on Drug Abuse (N.L.K.) and Award No. GM067193 from the National Institute of General Medical Sciences (N.L.K.). This research is based upon work supported by the National Science Foundation under grant No. ABI-1062432 to Indiana University (R.D.L.). This project is also funded, in part, with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views of policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, the National Center for Genome Analysis Support, Northwestern University, or Indiana University.

■ ABBREVIATIONS

AcN	acetonitrile
AEBSF	4-(2-aminoethyl)benzenesulfonyl fluoride
AGC	automatic gain control
ANOVA	analysis of variance
BCA	bicinchoninic acid
CV	coefficient of variation
dMS	differential mass spectrometry
DTT	dithiothreitol
FDR	false discovery rate
FT-ICR	Fourier transform ion cyclotron resonance
G-418	Geneticin, O-2-amino-2,7-didesoxy-D-glycero-*α*-D-gluco-heptopyranosyl-(1−4)-O-(3-desoxy-4-C-methyl-3-(methylamino)-*β*-L-arabinopyranosyl-(1−6))-D-streptamine
GELFrEE	gel elution liquid fractionation entrapment electrophoresis
HCD	higher energy collisional dissociation
IMT	intact mass tag
MS	mass spectrometry
MS1	intact/precursor scan
MS2	(or MS/MS), tandem mass spectrometry scan, fragmentation
nLC	nano liquid chromatography

QMT	quantitation mass target
SDS	sodium dodecyl sulfate
TMT	tandem mass tag
WT	wild type
YPD	yeast extract peptone dextrose
YPER	yeast protein extraction reagent

■ REFERENCES

(1) Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. *Anal. Chem.* **2011**, *83*, 6868−74.

(2) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.

(3) Smith, L. M.; Kelleher, N. L. *Nat. Methods* **2013**, *10*, 186−7.

(4) Kelleher, N. L. *Anal. Chem.* **2004**, *76*, 197A−203A.

(5) Reid, G. E.; McLuckey, S. A. *J. Mass Spectrom.* **2002**, *37*, 663−75.

(6) Rose, R. J.; Damoc, E.; Denisov, E.; Makarov, A.; Heck, A. J. R. *Nat. Methods* **2012**, *9*, 1084−6.

(7) Lowenthal, M. S.; Liang, Y.; Phinney, K. W.; Stein, S. E. *Anal. Chem.* **2014**, *86*, 551−8.

(8) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. *Mol. Cell Proteomics* **2005**, *4*, 1487−502.

(9) Wiese, S.; Reidegeld, K. A.; Meyer, H. E.; Warscheid, B. *Proteomics* **2007**, *7*, 340−350.

(10) Veenstra, T. D.; Martinovic, S.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 78−82.

(11) Ansong, C.; Wu, S.; Meng, D.; Liu, X. W.; Brewer, H. M.; Kaiser, B. L. D.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Pasa-Tolic, L. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 10153−10158.

(12) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Mol. Cell Proteomics* **2013**, *12*, 3465−73.

(13) Demirev, P. A.; Feldman, A. B.; Kowalski, P.; Lin, J. S. *Anal. Chem.* **2005**, *77*, 7455−7461.

(14) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. *Nature* **2011**, *480*, 254−8.

(15) Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L. *Anal. Chem.* **2006**, *78*, 4271−80.

(16) Dong, X. T.; Sumandea, C. A.; Chen, Y. C.; Garcia-Cazarin, M. L.; Zhang, J.; Balke, C. W.; Sumandea, M. P.; Ge, Y. *J. Biol. Chem.* **2012**, *287*, 848−857.

(17) Chamot-Rooke, J.; Mikaty, G.; Malosse, C.; Soyer, M.; Dumont, A.; Gault, J.; Imhaus, A. F.; Martin, P.; Trellet, M.; Clary, G.; Chafey, P.; Camoin, L.; Nilges, M.; Nassif, X.; Dumenil, G. *Science* **2011**, *331*, 778−782.

(18) Parks, B. A.; Jiang, L.; Thomas, P. M.; Wenger, C. D.; Roth, M. J.; Boyne, M. T., 2nd; Burke, P. V.; Kwast, K. E.; Kelleher, N. L. *Anal. Chem.* **2007**, *79*, 7984−91.

(19) Du, Y.; Parks, B. A.; Sohn, S.; Kwast, K. E.; Kelleher, N. L. *Anal. Chem.* **2006**, *78*, 686−94.

(20) Collier, T. S.; Sarkar, P.; Rao, B.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 879−889.

(21) Hung, C. W.; Tholey, A. *Anal. Chem.* **2012**, *84*, 161−170.

(22) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. *Genome Med.* **2013**, *5*, 53.

(23) Mazur, M. T.; Cardasis, H. L.; Spellman, D. S.; Liaw, A.; Yates, N. A.; Hendrickson, R. C. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 7728−7733.

(24) Meng, F. Y.; Wiener, M. C.; Sachs, J. R.; Burns, C.; Verma, P.; Paweletz, C. P.; Mazur, M. T.; Deyanova, E. G.; Yates, N. A.; Hendrickson, R. C. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 226−233.

(25) Wiener, M. C.; Sachs, J. R.; Deyanova, E. G.; Yates, N. A. *Anal. Chem.* **2004**, *76*, 6085−6096.

(26) Taylor, S. W.; Andon, N. L.; Bilakovics, J. M.; Lowe, C.; Hanley, M. R.; Pittner, R.; Ghosh, S. S. *J. Proteome Res.* **2006**, *5*, 1776−84.

(27) Ghaemmaghami, S.; Huh, W.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737−741.

(28) Hu, Y. H.; Rolfs, A.; Bhullar, B.; Murthy, T. V. S.; Zhu, C.; Berger, M. F.; Camargo, A. A.; Kelley, F.; McCarron, S.; Jepson, D.; Richardson, A.; Raphael, J.; Moreira, D.; Taycher, E.; Zuo, D. M.; Mohr, S.; Kane, M. F.; Williamson, J.; Simpson, A.; Bulyk, M. L.; Harlow, E.; Marsischky, G.; Kolodner, R. D.; LaBaer, J. *Genome Res.* **2007**, *17*, 536−543.

(29) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol. Cell Proteomics* **2013**, in press.

(30) Baudin, A.; Ozierkalogeropoulos, O.; Denouel, A.; Lacroute, F.; Cullin, C. *Nucleic Acids Res.* **1993**, *21*, 3329−3330.

(31) Jiang, L. *Analyzing Post Translational Modifications on Yeast Core Histones Using Fourier Transform Mass Spectrometry*. Dissertation/Thesis, University of Illinois at Urbana-Champaign, Ann Arbor, MI, 2008.

(32) Pile, L. A.; Schlag, E. M.; Wassarman, D. A. *Mol. Cell. Biol.* **2002**, *22*, 4965−76.

(33) Rundlett, S. E.; Carmen, A. A.; Kobayashi, R.; Bavykin, S.; Turner, B. M.; Grunstein, M. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 14503−14508.

(34) Henriksen, P.; Wagner, S. A.; Weinert, B. T.; Sharma, S.; Bacinskaja, G.; Rehman, M.; Juffer, A. H.; Walther, T. C.; Lisby, M.; Choudhary, C. *Mol. Cell Proteomics* **2012**, *11*, 1510−22.

(35) Wessel, D.; Flugge, U. I. *Anal. Biochem.* **1984**, *138*, 141−143.

(36) Vellaichamy, A.; Tran, J. C.; Catherman, A. D.; Lee, J. E.; Kellie, J. F.; Sweet, S. M.; Zamdborg, L.; Thomas, P. M.; Ahlf, D. R.; Durbin, K. R.; Valaskovic, G. A.; Kelleher, N. L. *Anal. Chem.* **2010**, *82*, 1234−44.

(37) Ahlf, D. R.; Compton, P. D.; Tran, J. C.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. *J. Proteome Res.* **2012**, *11*, 4308−14.

(38) Leduc, R. D.; Kelleher, N. L. *Curr. Protoc Bioinformatics* **2007**, Chapter 13, Unit 13 6; DOI: 10.1002/0471250953.bi1306s19.

(39) 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. *Anal. Chem.* **2012**, *84*, 209−15.

(40) LeDuc, R. D.; Boyne, M. T., 2nd; Townsend, R. R.; Bose, R. In *RECOMB Satellite Conference on Computational Proteomics 2010*; University of California: San Diego, CA, 2010.

(41) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513−23.

(42) Green, R.; Lesage, G.; Sdicu, A. M.; Menard, P.; Bussey, H. *Microbiology* **2003**, *149*, 2487−2499.

(43) Bommer, U. A.; Heng, C.; Perrin, A.; Dash, P.; Lobov, S.; Elia, A.; Clemens, M. J. *Oncogene* **2010**, *29*, 763−73.

(44) Posas, F.; Chambers, J. R.; Heyman, J. A.; Hoeffler, J. P.; de Nadal, E.; Arino, J. *J. Biol. Chem.* **2000**, *275*, 17249−55.

(45) Li, M.; Jiang, L.; Kelleher, N. L. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2009**, *877*, 3885−92.