HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2017 March 04.

Published in final edited form as:

J Proteome Res. 2016 March 4; 15(3): 976–982. doi:10.1021/acs.jproteome.5b00997.

Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa

Kenneth R. Durbin[†], **Luca Fornelli**[†], **Ryan T. Fellers**[†], **Peter F. Doubleday**[†], **Masashi Narita**[‡], and **Neil L. Kelleher**[*,†]

[†]Departments of Chemistry and Molecular Biosciences, Northwestern University, 2170 Campus Drive, Evanston, Illinois 60208, United States

[‡]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge CB2 0RE, U.K.

Abstract

Top-down proteomics is capable of identifying and quantitating unique proteoforms through the analysis of intact proteins. We extended the coverage of the label-free technique, achieving differential analysis of whole proteins <30 kDa from the proteomes of growing and senescent human fibroblasts. By integrating improved control software with more instrument time allocated for quantitation of intact ions, we were able to collect protein data between the two cell states, confidently comparing 1577 proteoform levels. To then identify and characterize proteoforms, our advanced acquisition software, named Autopilot, employed enhanced identification efficiency in identifying 1180 unique Swiss-Prot accession numbers at 1% false-discovery rate. This coverage of the low mass proteome is equivalent to the largest previously reported but was accomplished in 23% of the total acquisition time. By maximizing both the number of quantified proteoforms and their identification rate in an integrated software environment, this work significantly advances proteoform-resolved analyses of complex systems.

[*]Corresponding Author: n-kelleher@northwestern.edu. Phone: 847-467-4362. Fax: 847-467-3276..

ASSOCIATED CONTENT

Supporting Information

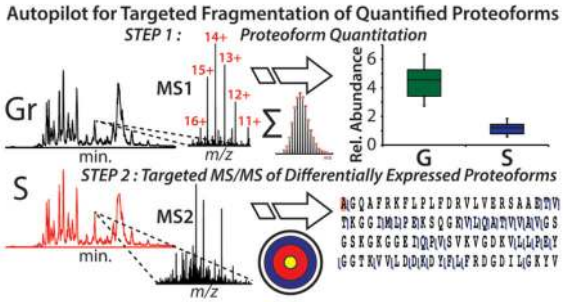
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00997. Percentage of time spent on MS^1 and MS^2 from a Top 2 data-dependent method; high level view of Autopilot workflow; top-down label-free quantitation of nuclei and cytoplasm from growing and oncogene-induced senescence; Western blot for ERH expression in growing and senescent fibroblasts; quantitation of different HMGA1 proteoforms represented with box and whisker plots; a silver stained SDS-PAGE gel to visualize each 10% GELFrEE fraction from growing and senescent lysates; general and detailed GO analyses for specific localization of proteins identified by Autopilot; higher sequence coverage obtained for proteolytically cleaved transgelin versus the full-length form; a "no enzyme" search in ProSight (called a Gene-Restricted BioMarker, or GRBM search) identified coatomer subunit ζ-1; internal fragmentation of ATP synthase subunit ε and results of GRBM searches (PDF)

List of QMTs differentially expressed in the cytoplasmic fraction of growing versus senescent fibroblasts, with related fold change (PDF)

List of QMTs differentially expressed in the nuclear fraction of growing versus senescent fibroblasts, with related fold change (PDF)

List of total proteoforms identified at 5% and 1% FDR; list of related unique accession numbers (XLSX)

The authors declare no competing financial interest.



Keywords

top-down proteomics; proteoform; cellular senescence; GELFrEE; Fourier transform mass spectrometry; label-free; quantitative proteomics; differential mass spectrometry

INTRODUCTION

In the recent past, top-down mass spectrometry has been relegated to analysis of individual intact proteins or simple mixtures. Meanwhile, complex mixtures were digested and analyzed by more established methods such as bottom-up proteomics. Even after pairing with liquid chromatography, top-down mass spectrometry only identified the highest abundance proteins with confidence.[1] However, an overhaul of the top-down proteomics (TDP) platform elevated the overall technique to a position where high numbers of identifications are now possible.[2] Recent advances in TDP can largely be attributed to improved protein separations, faster mass spectrometers, and computing solutions capable of processing hundreds of data files.[3–5]

While TDP can be operated in a high throughput mode to achieve the largest possible proteome coverage, the rate of identifying new proteins and proteoforms[6] after initial rounds of data collection reduces sharply.[5] The operation of current mass spectrometers has an inherent role in this TDP limitation because acquisition is not directed based on previously collected information. Recent instrumental improvements have led to faster scan times to boost peptide identifications, but this has not solved the challenge of obtaining tandem MS data of whole proteins over a broad dynamic range.[7] Electrospray of denatured, intact proteins produces a multitude of charge states that hampers straightforward approaches to detect and characterize low abundance proteoforms. Also, longer scan times and high target ion requirements of whole proteins limit the speed that mass spectra can be acquired. For these reasons, protein identification and proteoform characterization in TDP suffer from a dynamic range challenge where the same highly abundant species are repeatedly fragmented. Mass (not *m*/*z*) based inclusion and exclusion lists, especially when integrated into the acquisition software, are valuable to facilitate targeted analysis of lower intensity proteins over the course of a multiweek project. Through exclusion of previously acquired proteins, different charge states of the same protein can be avoided, and the instrument can focus on new protein signals. However, an exclusion list is not sufficient to reveal low abundance proteoforms. While instruments can make very sensitive measurements of intact proteins and their fragment ions, they must be directed to do so.[8] Acquisition software

supplied by instrument vendors is largely designed for bottom-up workflows and therefore does not contain capabilities to obtain tandem mass spectra on low abundance signals in a proteoform-resolved fashion. Instead, acquisition modes such as data-dependent, data-independent,[9] MS^E,[10] and others including decision trees[11] are popular methods for peptide data acquisition. Meanwhile, significantly less focus has been placed toward automating whole protein tandem MS by online[8,12] or offline analysis.[13] With all of these considerations in mind, a software package named Autopilot was recently developed with project-wide target/exclusion lists, real-time database searching, and optimized fragmentation during individual LC–MS runs.[8]

With intact protein identification now robust and headed toward "deep" coverage of the proteome <30 kDa, the move toward quantitative operation is a clear next step in the evolution of TDP. The combination of high-quality differential MS that is proteoform-resolved now offers a new view of molecular mechanisms and markers of disease without "blurring" the proteomic picture through digestion. Initial forays into top-down quantitation (TDQ) in a label-free format for differential analyses have shown promise.[14–16] Our group reported changes to over 100 yeast proteoforms using a platform for TDQ based on a label-free approach.[17] Additionally, the release around the same time of other comparative studies on the human saliva proteome,[18] apolipoproteins,[14] and *Salmonella typherium*[5] signals that differential TDQ MS will be viable to deploy in the near-term future. Most recently, a study was released detailing TDQ of 982 proteoforms from patient-derived breast tumor xenographs.[19]

We sought a cell-based system with pervasive phenotypic and morphological changes on which to benchmark our qualitative and differential top-down capabilities. Of interest here are human IMR90 fibroblasts, which undergo oncogene-induced senescence (OIS) upon induction of a constitutively active form of H-Ras.[20] While cellular senescence is a permanent cell cycle arrest typically associated with aging,[21] the phenomenon can also arise from various stressors. Cell cycle arrest upon oncogene introduction is a form of inherent resistance toward tumorigenesis in many cell types and, indeed, the onset of senescence in cells antagonizes tumorigenesis.[22] The senescence program makes cells enlarge and flatten, with large scale heterochromatin formation in the nucleus accompanied by metabolic reprogramming of primary metabolism as reported by us and another group in 2013.[23,24] OIS is initiated by complex signaling pathways that lead to downstream changes in transcription and proteomic alterations,[25–27] generating interest in determining the protein and modification dynamics operative in the senescence program.

To interrogate proteoform-level dynamics in cellular senescence, Autopilot[8] was improved and deployed to study OIS more efficiently than before in part by linking quantitative, then qualitative proteoform analyses. Here, Autopilot produced the largest quantitative study by TDP to date, exceeding the number of confidently quantitated proteoforms in the seminal paper by over 10-fold while making protein identification >4-fold more efficient than prior work on the system.[2]

MATERIALS AND METHODS

Cell Culture

Primary IMR90 human fibroblasts containing an ER:Ras fusion construct were grown adherently in Dulbecco's modified Eagle's medium (Sigma, St. Louis, MO) supplemented with 10% fetal bovine serum and 1% penicillin–streptomycin. Cells were treated with 100 nM 4-hydroxytamoxifen (4-OHT) to induce H–RasV12 expression and thereby senescence.[28] Adoption of OIS occurred after 7 days of Ras expression and was confirmed by a senescence-associated β-galactosidase staining protocol described previously.[29] Typically, 80–90% of cells exhibited senescent markers after 7 days of treatment with 4-OHT (Figure 1).

Sample Preparation

IMR90 cells were trypsinized and centrifuged at 270 × *g* for 5 min to pellet cells. The pellet was washed once with phosphate buffered saline and centrifuged again. The cells were differentially centrifuged into cytoplasmic and nuclear fractions.[30] Briefly, cells were resuspended in a buffer composed of 250 mM sucrose, 10 mM Tris-HCl pH 7.4, and 0.1 mM EGTA followed by homogenization with ~100 strokes of a Teflon homogenizer. The mixture was centrifuged at 2500 × *g* for 15 min to remove nuclei. The supernatant after centrifugation contained the cytoplasmic proteins. Fractions were lysed in 10 mM Tris-HCl pH 7.8, 4% SDS, 1 mM dithiothreitol, 10 mM sodium butyrate, and a cocktail of protease inhibitors (Thermo Pierce, Rockford, IL). The samples were vortexed and boiled for 10 min.

Lysed samples were acetone precipitated and resuspended in 1% SDS. Approximately 300 μg of protein was loaded per lane of a GELFREE 8100 Fractionation System (Expedeon, Harston, Cambridgeshire, UK). For high-throughput TDP, 10% GELFREE 8100 columns were used for molecular weight separation of nuclear and cytoplasmic fractions. The platform for label-free TDQ employed 8% GELFREE 8100 cartridges for collection of proteins <30 kDa into a single fraction as previously described.[17] GELFrEE (gel-eluted liquid-fraction entrapment electrophoresis) fractions were visualized by loading 1/20 of the volume of each fraction onto an SDS-PAGE gel and silver staining the gel.[31] Prior to MS analysis, all fractions were precipitated with a MeOH–$CHCl_3$–H_2O mixture to remove SDS detergent.[32] Precipitated protein was resuspended in 30–40 μL of Buffer A composed of 95% mass spectrometry-grade water, 4.8% acetonitrile, and 0.2% formic acid.

Liquid Chromatography–Mass Spectrometry Settings

Proteins were analyzed by nanoRPLC–MS/MS. For each TDP LC injection, 6 μL was loaded onto a trapping column (detailed below) with a Dionex Ultimate 3000 RSLC system (Thermo Fisher Scientific, Sunnyvale, CA) and washed with Buffer A for 10 min at 3 μL/min. The 90 min LC gradient used was previously described.[2] Autopilot-based identification used PLRP-S media (Agilent, Santa Clara, CA) packed in-house into 2 cm long × 150 μm inner diameter trapping columns coupled to 10 cm long × 75 μm inner diameter analytical columns. Alternatively, the TDQ platform utilized 2 cm Dionex Pepswift trapping columns and monolithic Thermo Dionex RP-4H analytical columns (100 μm ID × 50 cm long) at a flow rate of 1 μL/min and heated to 35 °C. Samples were loaded and washed for 3 min at 10

μL/min. The gradient was the same as for the TDP runs. Electrospray tips were packed with <1 cm of PLRP-S media to reduce bubble formation and promote stable ESI. An Orbitrap Elite (Thermo Fisher Scientific, San Jose, CA) collected all MS data. Xcalibur was the acquisition platform for TDQ quantitative scans where only MS^1 was performed. Autopilot acquired all TDP data. Parameters for Autopilot were as outlined previously.[8]

Top-down Proteomics Data Acquisition and Analysis

Data acquisition was performed by Autopilot, which has been presented in detail previously.[8] The precursor and fragmentation spectra were analyzed with Xtract (Thermo) for mass determination of intact proteins and fragment ions, respectively. Database searches by Autopilot utilized the "absolute mass mode" for searching MS^1 and MS^2 mass values against a "simple" database of 148 641 candidate proteoforms created from a human Uniprot flatfile. The simple database is created using the Database Manager tool in ProSightPC 3.0 from the "naked sequences" of all isoforms listed in the UniProt flatfile, allowing only for N-terminal acetylation and N-terminal Met cleavage as possible PTMs. The absolute mass mode search gathers all candidate proteoforms within a precursor mass tolerance for searching. The search tolerance for Autopilot was 250 Da. Data sets occasionally do not match any candidates within this small window size. Further offline processing employed larger, more computationally intensive searches in an effort to garner confident identifications for these unidentified species. To accomplish searches where large MS^1 window sizes are used (e.g., 2000 Da), a distributed version of ProSightPC 3.0 on a 168-node computing cluster was used. Cluster searches were performed as described previously,[2] with the lone deviation an addition of a Gene-Restricted Biomarker (GRBM) search after the absolute mass searches narrowed the number of possible candidates. The GRBM search mode seeks to characterize proteoforms araising from unknown proteolytic events using all subsequences derived from a single gene, previously identified through an absolute mass search. Performing a GRBM search can add an additional degree of specificity in the quality of characterization for previously identified gene products, usually localizing the termini of the truncated proteoform with single amino acid resolution. This addition of the GRBM search used no restriction of the MS^1 window size. The "complex" database contained 21 624 023 candidate proteoforms, generated using a far greater number of sources of protein variation curated in the Swiss-Prot database, with up to 14 PTMs (or known polymorphisms) for each isoform listed in an individual accession number. Instantaneous protein-level FDR calculations were performed as described previously using scrambled decoy databases and a procedure that assigns a *q*-value to each identification.[3]

Top-down Label-Free Quantitation (TDQ) Platform

In the workflow implemented here, the quantitated samples contained proteins grouped by size from ~5–30 kDa using the first fraction collected from a GELFrEE separation run with an 8% T gel. For differential analysis of the two fibroblast populations, three biological and 6–7 technical replicates were analyzed by quantitative nanoRPLC–MS in randomized order. The technical replicates obtained from a given population of cells (i.e., senescent or control) were analyzed with only precursor scans and no fragmentation. The applied resolution was 120 000 (at 400 *m/z*) with 15 μscans for each MS^1 scan. Following quantitative data collection, statistical analysis with a linear hierarchical model calculated proteoform fold-

differences, assigning variation to the technical process and scoring proteoform changes in the system.[33] The output of quantified proteoforms was visualized using a volcano plot representation. Results better than a 5% FDR cutoff were called confident quantitative mass targets (QMTs).[17] Follow-up LC–MS/MS runs directed by AUTOPILOT then performed proteoform identification. Identifications were linked to QMTs through 10 ppm intact mass tag searches.

Western Blot

Growing and senescent cells were washed with PBS and separately lysed in lysis buffer as detailed above. Crude lysate underwent one freeze/thaw cycle and was further disrupted by sonication. Protein concentration was normalized by BCA, samples were resolved by SDS-PAGE using a 4–20% gel (Bio-Rad), and protein was transferred to a PVDF membrane. The immunoblots used monoclonal antibodies against ERH (rabbit, Abcam EPR10830[B]) and alpha-Tubulin (mouse, Sigma, T9026) incubated overnight at 4 °C in ratios of 1:1000 and 1:2000, respectively. Blots were probed with HRP-conjugated antibodies to mouse (1:20 000) and rabbit (1:10 000) and visualized with the chemiluminescent reagent SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Scientific) using a Bio-Rad ChemiDoc XRS imaging system. All blots were performed in triplicate.

RESULTS AND DISCUSSION

Large-Scale Implementation of Top-down Quantitation in a Label-Free Format

Previous TDQ data were acquired with a mass spectrometry method where two of every three scans recorded fragmentation spectra for protein identification (i.e., a “data-dependent top two” method).[17] In LC–MS/MS runs using such acquisition logic, only 25% of instrument time is spent acquiring MS^1 scans on proteoforms (Supplemental Figure S1). To improve the quality of quantitative information derived from MS^1 scans, we removed tandem MS^2 scans entirely and implemented full scan-only LC–MS runs (Figure 1, top right). Subsequently, optimized MS^2 scans were performed on replicate samples by AUTOPILOT (Supplemental Figure S2) after data sets for label-free quantitation were collected and processed (Figure 1, bottom right). From replicate analyses of cytoplasmic extracts from human fibroblasts using “MS^1-only” runs, the levels of 1955 QMTs were determined after comparison of control growing cells to those undergoing OIS following Ras expression (Supplemental Figure S3A). Induction of senescence was confirmed after 7 days of treatment with 4-OHT by senescence associated β-galactosidase (SA-β-gal) positive staining (Figure 1, top left). Treated cells exhibited 87% SA-β-gal positive cells, while untreated cells displayed 10% positive cells. Overall, 1038 of the 1955 total QMTs (i.e., proteoforms) exceeded the arbitrary threshold of a 5% instantaneous FDR, meaning their *q-value* was <0.05 (see Figure 2A). The mass values of these 1,038 QMTs were used to match to known identifications in the same cells determined by tandem MS within a 10 ppm tolerance (Figure 2A, yellow circles; Supplemental Table S1). In all, 751 identifications were made from the pool of 1038 QMTs with a confidence level of 5% FDR or better.

Of the 751 identified proteoforms that were confidently quantified, one form of enhancer of rudimentary homologue (ERH) was downregulated 3.6-fold in senescent cells (Figure 2B

and Supplemental Figure S3A, right). ERH has previously been implicated in regulating nuclear gene expression and downregulation was shown to produce inverted KRas signatures.[34] As such, regulation of ERH may play a role in the senescence program brought on through oncogene expression. The downregulation of the ERH protein family has also been confirmed by Western blot (Supplemental Figure S4). Two other proteoforms with even larger differential abundance changes in senescence are highlighted in Figure 2. FAM107B isoform 1 (Figure 2C) displayed a nine-fold decrease in OIS, while guanine nucleotide-binding protein (Figure 2D) was upregulated six-fold; the metrics of confidence associated with these fold changes are reported and explained in the caption of Figure 2.

From the nuclear fraction of fibroblast cells, 1407 QMTs were determined to be differentially expressed after data analysis (Supplemental Figure S3B). Of these, 539 were found at, or better than, the 5% FDR confidence level (Supplemental Table S2). Cystatin-B, previously associated with senescence, was increased by ~2-fold (Supplemental Figure S3B, far right).[35] Additionally, 35 proteoforms of histone proteins H2A, H3, and H4 were quantified, with proteoforms from each of the three histone H3 isoforms independently tracked in a gene-specific fashion (i.e., H3.1, H3.2, and H3.3; see Supplemental Table S2). Along with histones, members of the HMG protein family are highly modified nuclear proteins. Acetylation, methylation, and phosphorylation are often present on histone and HMG proteins,[36] creating multiple proteoforms, which were identified and quantitated. Several proteoforms of HMGA1, a protein previously shown to promote senescence-associated heterochromatin foci formation, had changed expression abundance (Supplemental Figure S5).[37]

To our knowledge, the scale of this study represents the largest report of differential top-down proteomics run in discovery-mode.[17] Performing MS^2 for protein identification only after quantitative data were collected enabled four-fold more instrument time (compared to data-dependent acquisition) dedicated to MS^1 scans; this increased the number of MS^1 measurements for eluting species, improved signal-to-noise ratios (S/N), and translated into both a greater number of QMTs and improved metrics of their statistical confidence (i.e., lowered *q*-values). Additionally, the incorporation of the monolithic nanocapillary columns into our LC step provided high reproducibility, sharp peak shapes, and reduced carryover between runs compared to our last report.[17]

Autopilot Improves Efficiency of Qualitative Analysis by ~4-Fold

To maximize the efficiency of TDP (as defined by the number of LC–MS runs required to reach a certain level of proteome coverage), Autopilot was applied to GELFrEE fractions from prefractionated whole cell extracts. Both growing and 4-OHT treated IMR90 fibroblasts were divided into cytoplasmic and nuclear fractions by differential centrifugation. After each fraction was separated by molecular weight using GELFrEE, the fractions were composed of proteins in ~4–5 kDa windows (e.g., 10–15 kDa; see Supplemental Figure S6).[38] From a total of 98 LC–MS runs controlled by Autopilot and collected in less than 1 week of instrument time, we generated a total of 1599 protein identifications (defined as unique UniProt accession numbers at 5% FDR, see Supplemental Table S3 for all unique proteoforms identified). The subcellular localization, as predicted by

Gene Ontology, of the proteins identified is provided in Supplemental Figure S7A. The number of proteins localized to the nucleus and mitochondria was 497 and 312 proteins, respectively. The specific localization of proteins from the nucleus and mitochondria is shown in Supplemental Figure S7B,C. Only GELFrEE fractions with proteins under 30 kDa were injected, and all mass determinations were made by algorithmic interpretation of resolved isotopic distributions.

Initial database searches were performed in real-time by Autopilot during LC–MS data acquisition.[8] Overall, 1157 unique proteins were found by these online searches. A small, 250 Da search window was used for real-time searches with Autopilot to maintain pace with instrument acquisition. After runs were completed, absolute mass searches with far larger MS^1 windows for error tolerance found 1582 total unique proteins below a 5% FDR cutoff (P-Score $<6.1 \times 10^{-8}$), while gene-restricted Biomarker type searches ("GRBM") led to 701 total proteins below FDR cutoff (P-Score $<6.2 \times 10^{-12}$). For the 701 proteins identified by GRBM searching, 684 over-lapped with absolute mass identifications (Figure 3A). At 1% FDR, 1180 unique protein identifications were identified. From this set, 1148 proteins were found through absolute mass searches (596 unique), and 584 proteins were found by GRBM searches (32 unique). The number of unique protein identifications <30 kDa is nearly equivalent to the largest previous TD study (1598 unique human accession numbers <30 kDa at 5% FDR).[2] That study employed 423 LC–MS/MS runs to reach that level of coverage, whereas we achieved 100% of the number of identifications in under 25% of the number of comparable LC runs (Figure 3B). The ~4-times increase in efficiency can be attributed to the presence of project-wide exclusion lists, improved fragmentation of unidentified proteins, and SIM scans to detect low abundant species in Autopilot-based acquisition.[8]

The use of GRBM searches interrogates all subsequences of candidate proteins. GRBM searches were utilized to find proteoforms that arise from proteolytic clipping of fragments shed from larger proteins >30 kDa that would not be confidently identified by normal error-tolerant database queries. With this method of search, a truncated form of transgelin was discovered through GRBM searching, whereas originally only fragment ions from one terminus had been matched (Supplemental Figure S8). Several instances of GRBM search results led to the identification of proteins that did not clear the FDR cutoff when searched in absolute mass mode. In Supplemental Figure S9, coatomer subunit ζ-1 was identified after 11 y-ions were matched following removal of the two C-terminal residues of the protein. The additional matching fragment ions substantially improved the P-Score from 3.1×10^{-6} (which corresponds to a q-value not making the FDR cutoff for absolute mass searches) to 7.5×10^{-25} (well below the FDR cutoff for identification by GRBM-type searches). Also, several GRBM searches with better fragmentation coverage were found even though the detected intact mass was within a small error tolerance from the theoretical mass of the full protein. For example, ATP synthase subunit ε gave multiple GRBM hits with strong results (Supplemental Figure S10). These hits correspond to internal fragment ions, which can supply additional information to support proteoform characterization.[39,40]

CONCLUSION

Several improvements to data acquisition for whole protein LC–MS were demonstrated here in a human senescence model. The result was substantially increased efficiency for collection of differential data on thousands of proteoforms in discovery mode. Overall, one of the largest TDP studies to date was completed in four-fold less acquisition time than previously required, making TDP acquisition more viable in terms of instrument and personnel commitment. As robust platforms for label-free top-down quantitation continue their evolution, proteoform-level differences can now be measured with a quantifiable metric of confidence. The use of MS^1-only runs in this work, followed by Autopilot-directed MS^2 on subsequent injections (with HCD energy optimized in real time), sharply improves the entire process of TDQ and TDP.

What remains to be understood is the intrinsic value of "the proteoform" in basic and clinical research. One seeks to increase the effciency of translating MS-based proteomics data into clinical advance and functional insight. While the value proposition of proteoform-resolved measurement is cloudy at the moment, this work advances the ongoing shift in TDP from a qualitative to a quantitatively enabled experiment. Most recently, a noteworthy example of this shift was observable. In the Ntai et al. publication, patient-derived xenographs from different breast cancer subtypes were analyzed by TDQ to uncover post-translational modification patterns not observable by bottom-up proteomics.[19] With the recent technology advances, researchers can now maximize proteome coverage and proteoform characterization on large sets of quantitative proteoform-level data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Nos. GM067193 and GM108569 (N.L.K.), federal funds from the National Cancer Institute (Office of Cancer Clinical Proteomics Research) under Contract No. HHSN261200800001E, and the Office of Research at Northwestern University. Additional support was provided by the UIUC Center for Neuroproteomics on Cell to Cell Signaling (P30 DA018310) and the Robert H. Lurie Comprehensive Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.N. acknowledges the University of Cambridge, the Cancer Research UK Cambridge Institute Core Grant, and Hutchison Whampoa. Additionally, L.F. would like to acknowledge the Swiss National Science Foundation for support of an Early Postdoc Mobility fellowship. We would also like to thank Ioanna Ntai, Phil Compton, Paul Thomas, Bryan Early, and Joe Greer as well as the rest of the members of the Kelleher Research Group for their help with this work.

REFERENCES

(1). Parks BA, Jiang L, Thomas PM, Wenger CD, Roth MJ, Boyne MT II, Burke PV, Kwast KE, Kelleher NL. Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers. Anal. Chem. 2007; 79:7984–7991. [PubMed: 17915963]

(2). Catherman AD, Durbin KR, Ahlf DR, Early BP, Fellers RT, Tran JC, Thomas PM, Kelleher NL. Large-scale Top-down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence. Mol. Cell. Proteomics. 2013; 12:3465–3473. [PubMed: 24023390]

(3). Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SMM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM,

Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011; 480:254–8. [PubMed: 22037311]

(4). Ahlf DR, Compton PD, Tran JC, Early BP, Thomas PM, Kelleher NL. Evaluation of the Compact High-Field Orbitrap for Top-Down Proteomics of Human Cells. J. Proteome Res. 2012; 11:4308–4314. [PubMed: 22746247]

(5). Ansong C, Wu S, Meng D, Liu X, Brewer HM, Deatherage Kaiser BL, Nakayasu ES, Cort JR, Pevzner P, Smith RD, Heffron F, Adkins JN, Pasa-Tolic L. Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:10153–10158. [PubMed: 23720318]

(6). Smith LM, Kelleher NL, Proteomics CTD, et al. Proteoform: a single term describing protein complexity. Nat. Methods. 2013; 10:186–187. [PubMed: 23443629]

(7). Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ. The One Hour Yeast Proteome. Mol. Cell. Proteomics. 2014; 13:339–347. [PubMed: 24143002]

(8). Durbin KR, Fellers RT, Ntai L, Kelleher NL, Compton PD. Autopilot: An Online Data Acquisition Control System for the Enhanced High-Throughput Characterization of Intact Proteins. Anal. Chem. 2014; 86:1485–1492. [PubMed: 24400813]

(9). Gillet LC, Navarro P, Tate S, Roest H, Selevsek N, Reiter L, Bonner R, Aebersold R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Mol. Cell. Proteomics. 2012; 11 O111.016717.

(10). Bond NJ, Shliaha PV, Lilley KS, Gatto L. Improving Qualitative and Quantitative Performance for MSE-based Label-free Proteomics. J. Proteome Res. 2013; 12:2340–2353. [PubMed: 23510225]

(11). Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods. 2008; 5:959–964. [PubMed: 18931669]

(12). Johnson JR, Meng FY, Forbes AJ, Cargile BJ, Kelleher NL. Fourier-transform mass spectrometry for automated fragmentation and identification of 5–20 kDa proteins in mixtures. Electrophoresis. 2002; 23:3217–3223. [PubMed: 12298093]

(13). Zhang H, Ge Y. Comprehensive Analysis of Protein Modifications by Top-Down Mass Spectrometry. Circ.: Cardiovasc. Genet. 2011; 4:711. [PubMed: 22187450]

(14). Mazur MT, Cardasis HL, Spellman DS, Liaw A, Yates NA, Hendrickson RC. Quantitative analysis of intact apolipoproteins in human HDL by top-down differential mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 2010; 107:7728–7733. [PubMed: 20388904]

(15). Taylor SW, Andon NL, Bilakovics JM, Lowe C, Hanley MR, Pittner R, Ghosh SS. Efficient high-throughput discovery of large peptidic hormones and biomarkers. J. Proteome Res. 2006; 5:1776–1784. [PubMed: 16823986]

(16). Meng FY, Wiener MC, Sachs JR, Burns C, Verma P, Paweletz CP, Mazur MT, Deyanova EG, Yates NA, Hendrickson RC. Quantitative analysis of complex peptide mixtures using FTMS and differential mass spectrometry. J. Am. Soc. Mass Spectrom. 2007; 18:226–233. [PubMed: 17070068]

(17). Ntai I, Kim K, Fellers RT, Skinner OS, Smith AD, Early BP, Savaryn JP, LeDuc RD, Thomas PM, Kelleher NL. Applying Label-Free Quantitation to Top Down Proteomics. Anal. Chem. 2014; 86:4961–4968. [PubMed: 24807621]

(18). Wu S, Brown JN, Tolic N, Meng D, Liu X, Zhang H, Zhao R, Moore RJ, Pevzner P, Smith RD, Pasa-Tolic L. Quantitative analysis of human salivary gland-derived intact proteome using top-down mass spectrometry. Proteomics. 2014; 14:1211–1222. [PubMed: 24591407]

(19). Ntai I, LeDuc RD, Fellers RT, Erdmann-Gilmore P, Davies SR, Rumsey J, Early BP, Thomas PM, Li S, Compton PD, Ellis MJ, Ruggles KV, Fenyo D, Boja ES, Rodriguez H, Townsend RR, Kelleher NL. Integrated Bottom-up and Top-down Proteomics of Patient-derived Breast Tumor Xenografts. Mol. Cell. Proteomics. 2016; 15:45. [PubMed: 26503891]

(20). Serrano M, Lin AW, McCurrach ME, Beach D, Lowe SW. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16(INK4a). Cell. 1997; 88:593–602. [PubMed: 9054499]

(21). Campisi J. The biology of replicative senescence. Eur. J. Cancer. 1997; 33:703–709. [PubMed: 9282108]

(22). Bennecke M, Kriegl L, Bajbouj M, Retzlaff K, Robine S, Jung A, Arkan MC, Kirchner T, Greten FR. Ink4a/Arf and Oncogene-Induced Senescence Prevent Tumor Progression during Alternative Colorectal Tumorigenesis. Cancer Cell. 2010; 18:135–146. [PubMed: 20708155]

(23). Kaplon J, Zheng L, Meissl K, Chaneton B, Selivanov VA, Mackay G, van der Burg SH, Verdegaal EM, Cascante M, Shlomi T, Gottlieb E, Peeper DS. A key role for mitochondrial gatekeeper pyruvate dehydrogenase in oncogene-induced senescence. Nature. 2013; 498:109–112. [PubMed: 23685455]

(24). Li M, Durbin KR, Sweet SM, Tipton JD, Zheng Y, Kelleher NL. Oncogene-induced cellular senescence elicits an anti-Warburg effect. Proteomics. 2013; 13:2585–2596. [PubMed: 23798001]

(25). Lanigan F, Geraghty JG, Bracken AP. Transcriptional regulation of cellular senescence. Oncogene. 2011; 30:2901–2911. [PubMed: 21383691]

(26). Courtois-Cox S, Jones SL, Cichowski K. Many roads lead to oncogene-induced senescence. Oncogene. 2008; 27:2801–2809. [PubMed: 18193093]

(27). de Graaf EL, Kaplon J, Zhou H, Heck AJR, Peeper DS, Altelaar AFM. Phosphoproteome Dynamics in Onset and Maintenance of Oncogene-induced Senescence. Mol. Cell. Proteomics. 2014; 13:2089–2100. [PubMed: 24961811]

(28). Chandra T, Kirschner K, Thuret J-Y, Pope BD, Ryba T, Newman S, Ahmed K, Samarajiwa SA, Salama R, Carroll T, Stark R, Janky R. s. Narita M, Xue L, Chicas A, Nunez S, Janknecht R, Hayashi-Takanaka Y, Wilson MD, Marshall A, Odom DT, Babu MM, Bazett-Jones DP, Tavare S, Edwards PAW, Lowe SW, Kimura H, Gilbert DM, Narita M. Independence of Repressive Histone Marks and Chromatin Compaction during Senescent Heterochromatic Layer Formation. Mol. Cell. 2012; 47:203–214. [PubMed: 22795131]

(29). Debacq-Chainiaux F, Erusalimsky JD, Campisi J, Toussaint O. Protocols to detect senescence-associated beta-galactosidase (SA-beta gal) activity, a biomarker of senescent cells in culture and in vivo. Nat. Protoc. 2009; 4:1798–1806. [PubMed: 20010931]

(30). Catherman AD, Li M, Tran JC, Durbin KR, Compton PD, Early BP, Thomas PM, Kelleher NL. Top Down Proteomics of Human Membrane Proteins from Enriched Mitochondrial Fractions. Anal. Chem. 2013; 85:1880–1888. [PubMed: 23305238]

(31). Shevchenko A, Wilm M, Vorm O, Mann M. Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. Anal. Chem. 1996; 68:850–858. [PubMed: 8779443]

(32). Wessel D, Flugge UI. A method for the quantitative recovery of proteins in dilute-solution in the presence of detergents and lipids. Anal. Biochem. 1984; 138:141–143. [PubMed: 6731838]

(33). Ntai I, Kim K, Fellers RT, Skinner OS, Smith A. D. t. Early BP, Savaryn JP, LeDuc RD, Thomas PM, Kelleher NL. Applying label-free quantitation to top down proteomics. Anal. Chem. 2014; 86:4961–4968. [PubMed: 24807621]

(34). Weng MT, Lee JH, Wei SC, Li QN, Shahamatdar S, Hsu D, Schetter AJ, Swatkoski S, Mannan P, Garfield S, Gucek M, Kim MKH, Annunziata CM, Creighton CJ, Emanuele MJ, Harris CC, Sheu JC, Giaccone G, Luo J. Evolutionarily conserved protein ERH controls CENP-E mRNA splicing and is required for the survival of KRAS mutant cancer cells. Proc. Natl. Acad. Sci. U. S. A. 2012; 109:E3659–E3667. [PubMed: 23236152]

(35). Dean JP, Nelson PS. Profiling influences of senescent and aged fibroblasts on prostate carcinogenesis. Br. J. Cancer. 2008; 98:245–249. [PubMed: 18182995]

(36). Zhang Q, Wang Y. HMG modifications and nuclear function. Biochim. Biophys. Acta, Gene Regul. Mech. 2010; 1799:28–36.

(37). Narita M, Narita M, Krizhanovsky V, Nunez S, Chicas A, Hearn SA, Myers MP, Lowe SW. A novel role for high-mobility group a proteins in cellular senescence and heterochromatin formation. Cell. 2006; 126:503–514. [PubMed: 16901784]

(38). Tran JC, Doucette AA. Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. Anal. Chem. 2008; 80:1568–1573. [PubMed: 18229945]

(39). Savaryn JP, Skinner OS, Fornelli L, Fellers RT, Compton PD, Terhune SS, Abecassis MM, Kelleher NL. Targeted analysis of recombinant NF kappa B (RelA/p65) by denaturing and native top down mass spectrometry. J. Proteomics. 2015 DOI:10.1016/j.jprot.2015.04.025.

(40). Durbin KR, Skinner OS, Fellers RT, Kelleher NK. Analyzing internal fragmentation of electrosprayed ubiquitin ions during beam-type collisional dissociation. J. Am. Soc. Mass Spectrom. 2015; 26:782–787. [PubMed: 25716753]

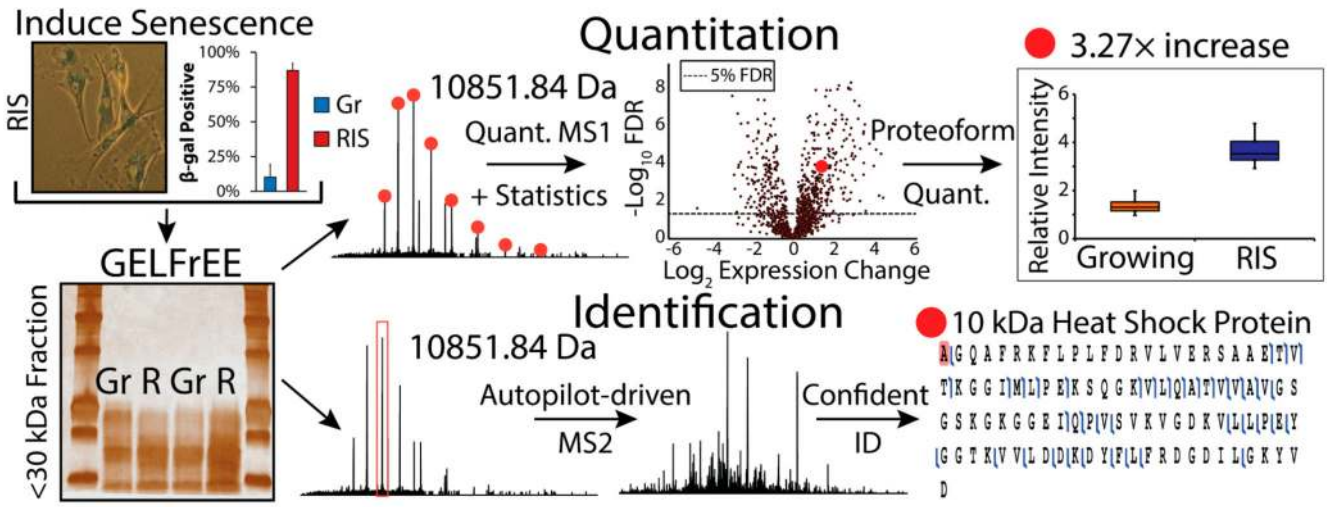


Figure 1.

The process of TDQ and intelligent identification of differentially modified proteoforms from senescent versus control cells. Senescence was induced in growing (Gr) IMR90 fibroblasts by oncogenic Ras expression and verified by positive SA-β-gal staining (top left). Error bars are standard deviation. After positive senescent cells were obtained, growing and Ras-induced senescent (R or RIS) cells were separated by molecular weight using GELFrEE (visualized using a silver-stained slab gel, bottom left). Fractions containing <30 kDa proteins were analyzed by Fourier transform mass spectrometry. (Top) First, MS^1-only spectra were acquired and processed through a linear statistical model. Fold-changes in proteoform abundance between senescent and growing cells were displayed in a volcano plot (upper right). (Bottom) Autopilot acquisition generates confident identifications of previously quantitated proteoforms by tandem MS. Identification-centric runs were performed only after all proteoform quantitation was complete.

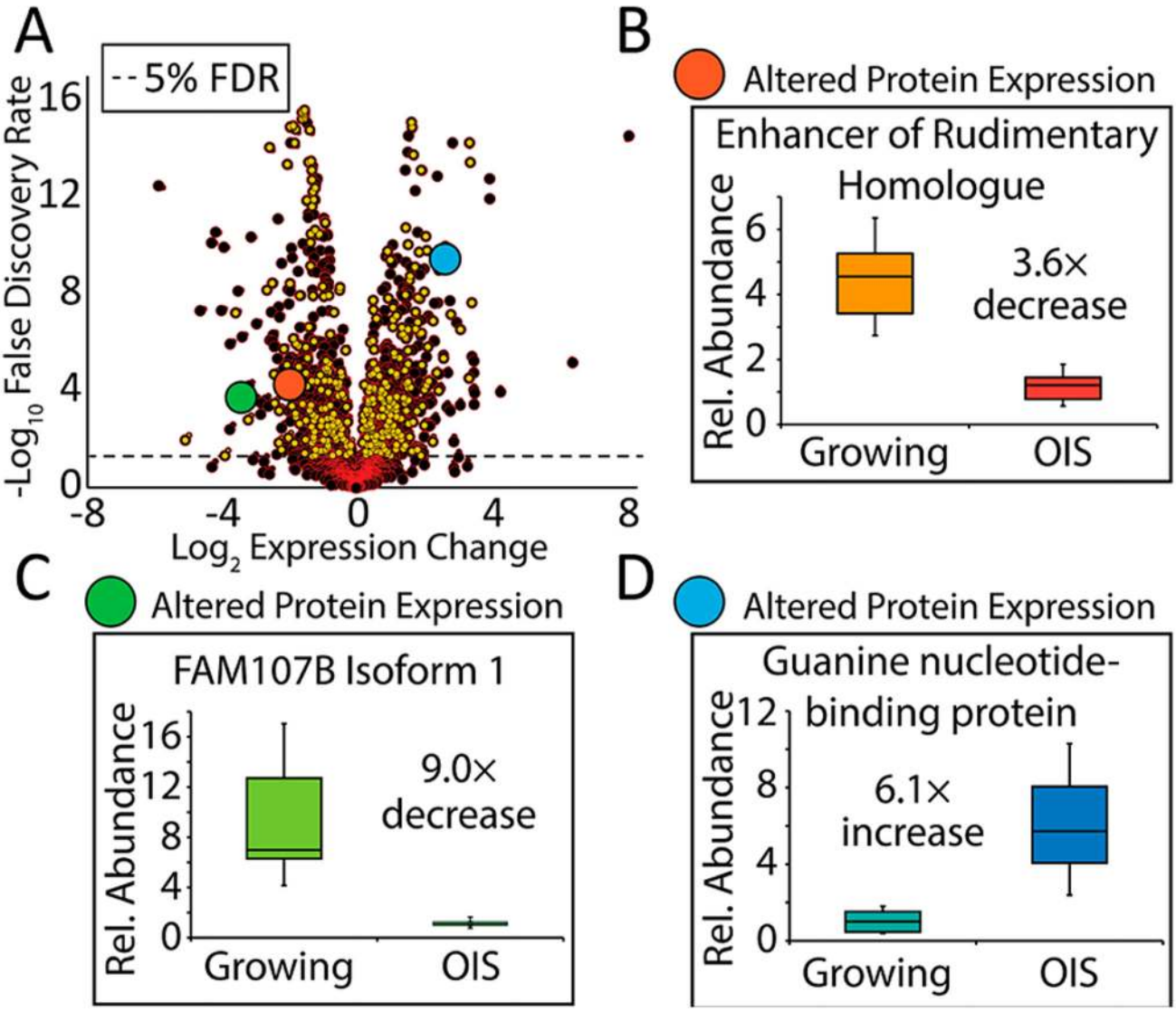


Figure 2.

Top-down quantitation yields changes in abundance of 1038 cytoplasmic proteoforms. (A) Differences in proteoform abundance are mapped according to their $\log_2$ fold change (*x*-axis) and the $-\log_{10}$ of a confidence metric (*y*-axis) best described as an instantaneous FDR (also called a *q*-value). The dotted line indicates the arbitrary 5% FDR threshold (i.e., *q*-values of 0.05 and below). Negative values on the *x*-axis signify a decrease in the level of the proteoform in senescence, while positive values indicate an increase. Each QMT/proteoform is indicated with a dot, with the identified QMTs denoted by a yellow color. The three large circles are selected examples highlighted in panels B–D. (B) A proteoform of the enhancer of rudimentary homologue (P84090) was downregulated by 3.6-fold in OIS compared to growing cells; the *q*-value associated with this observation was 0.00005, which converts via $-\log_{10}$ into 4.3. (C) One of the largest decreases in senescence was nine-fold, observed with a *q*-value of 0.0016 for Isoform 1 of FAM107B (large green circle; Q9H098).

(D) Guanine nucleotide-binding protein subunit gamma-5 (GBG5; P63218) was upregulated 6.1-fold in senescence with a q-value of 5.0×10^{-10} (or converted via $-\log_{10}$, 9.3). In the box and whisker plots, the first and third quartiles are the ends of the boxes with the median included. The whisker demarcate the minimum and maximum data points for the proteoform.

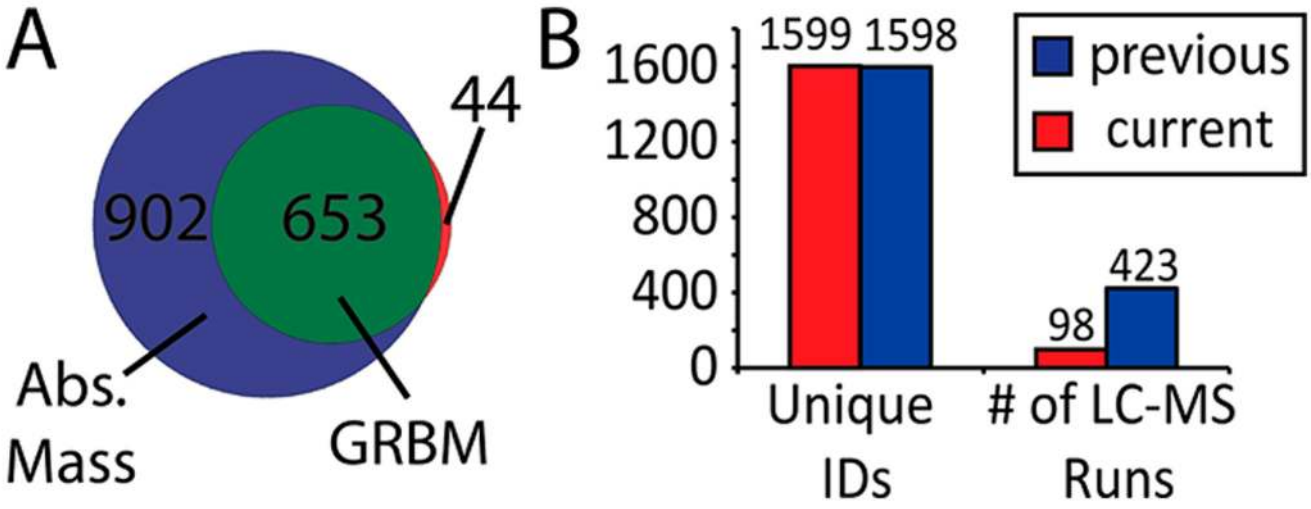


Figure 3.

Advanced Autopilot acquisition confidently identified 1599 unique Uniprot accession numbers from one subcellular fractionation experiment. (A) Multitiered searching was utilized to maximize total proteome coverage. Absolute mass searches (windowed search type based around intact mass) yielded 1555 identifications <30 kDa at a 5% FDR threshold. Through additional GRBM searches aimed at improving proteoform characterization, 44 new proteins were identified, while 653 were shared. The main goal of GRMB searching was to improve proteoform characterization. (B) A comparison of the number of confident identifications found under 30 kDa between this study and a previous one is shown.[2] Additionally, a comparison of the number of LC–MS runs needed to achieve each set of results reveals the same number of protein identifications were obtained in 1/4 of the time.