Large-scale Top-down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence*[S]

Adam D. Catherman‡, Kenneth R. Durbin‡, Dorothy R. Ahlf‡, Bryan P. Early‡, Ryan T. Fellers‡, John C. Tran‡, Paul M. Thomas‡, and Neil L. Kelleher‡§

Top-down proteomics is emerging as a viable method for the routine identification of hundreds to thousands of proteins. In this work we report the largest top-down study to date, with the identification of 1,220 proteins from the transformed human cell line H1299 at a false discovery rate of 1%. Multiple separation strategies were utilized, including the focused isolation of mitochondria, resulting in significantly improved proteome coverage relative to previous work. In all, 347 mitochondrial proteins were identified, including ~50% of the mitochondrial proteome below 30 kDa and over 75% of the subunits constituting the large complexes of oxidative phosphorylation. Three hundred of the identified proteins were found to be integral membrane proteins containing between 1 and 12 transmembrane helices, requiring no specific enrichment or modified LC-MS parameters. Over 5,000 proteoforms were observed, many harboring post-translational modifications, including over a dozen proteins containing lipid anchors (some previously unknown) and many others with phosphorylation and methylation modifications. Comparison between untreated and senescent H1299 cells revealed several changes to the proteome, including the hyperphosphorylation of HMGA2. This work illustrates the burgeoning ability of top-down proteomics to characterize large numbers of intact proteoforms in a high-throughput fashion. ***Molecular & Cellular Proteomics 12: 10.1074/mcp.M113.030114, 3465–3473, 2013.***

Although traditional bottom-up approaches to mass-spectrometry-based proteomics are capable of identifying thousands of protein groups from a complex mixture, proteolytic digestion can result in the loss of information pertaining to post-translational modifications and sequence variants (1, 2). The recent implementation of top-down proteomics in a high-throughput format using either Fourier transform ion cyclotron resonance (3–5) or Orbitrap instruments (6, 7) has shown an increasing scale of applicability while preserving information on combinatorial modifications and highly related sequence variants. For example, the identification of over 500 bacterial proteins helped researchers find covalent switches on cysteines (7), and over 1,000 proteins were identified from human cells (3). Such advances have driven the detection of whole protein forms, now simply called proteoforms (8), with several laboratories now seeking to tie these to specific functions in cell and disease biology (9–11).

The term "proteoform" denotes a specific primary structure of an intact protein molecule that arises from a specific gene and refers to a precise combination of genetic variation, splice variants, and post-translational modifications. Whereas special attention is required in order to accomplish gene- and variant-specific identifications via the bottom-up approach, top-down proteomics routinely links proteins to specific genes without the problem of protein inference. However, the fully automated characterization of whole proteoforms still represents a significant challenge in the field. Another major challenge is to extend the top-down approach to the study of whole integral membrane proteins, whose hydrophobicity can often limit their analysis via LC-MS (5, 12–16). Though integral membrane proteins are often difficult to solubilize, the long stretches of sequence information provided from fragmentation of their transmembrane domains in the gas phase can actually aid in their identification (5, 13).

In parallel to the early days of bottom-up proteomics a decade ago (17–21), in this work we brought the latest methods for top-down proteomics into combination with subcellular fractionation and cellular treatments to expand coverage of the human proteome. We utilized multiple dimensions of separation and an Orbitrap Elite mass spectrometer to achieve large-scale interrogation of intact proteins derived from H1299 cells. For this focus issue on post-translational modifications, we report this summary of findings from the largest implementation of top-down proteomics to date, which resulted in the identification of 1,220 proteins and thousands more proteoforms. We also applied the platform to H1299 cells induced into senescence by treatment with the DNA-damaging agent camptothecin.

From the ‡Departments of Chemistry and Molecular Biosciences, the Chemistry of Life Processes Institute, the Proteomics Center of Excellence, and the Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Evanston, Illinois 60208

Received April 16, 2013, and in revised form, September 5, 2013
Published, MCP Papers in Press, September 10, 2013, DOI 10.1074/mcp.M113.030114

EXPERIMENTAL PROCEDURES

Cell Culture and Treatment—NCI-H1299 cells (ATCC CRL 5803) were grown in Dulbecco's modified Eagle's medium (Sigma, St. Louis, MO) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin. To induce senescence, cells were treated with 25 nM camptothecin for 24 h and then allowed to recover for 4 days in normal media. We detected senescent cells by staining for β-galactosidase expression and conducting morphological examination under a light microscope (22). Both untreated and senescent cells were treated with 0.05% trypsin-EDTA solution (Invitrogen, Carlsbad, CA) and harvested via centrifugation at 500 $\times$ *g* for 3 min. The resulting cell pellets (~5 $\times$ 10^7 cells) were washed twice with phosphate-buffered saline before being flash-frozen in liquid nitrogen.

Subcellular Fractionation—Whole cell lysate was prepared by boiling the cell pellet in 150 mM Tris-HCl, 10 mM DTT, 4% SDS, pH 7.5, and then centrifuging to remove cellular debris. Alternatively, subcellular fractions were prepared by suspending the cells in ~20 ml of sucrose-Tris buffer (250 mM sucrose, 10 mM Tris-HCl, pH 7.4, 0.1 mM EGTA, and inhibitors (1% protease and phosphatase inhibitors, 10 mM sodium butyrate, 1 mM DTT)). The cells were lysed using ~50 strokes of a Teflon homogenizer. The lysate was centrifuged at 500 $\times$ *g* for 5 min to pellet unbroken cells. The cell lysis procedure was repeated until only a small pellet remained. The lysate was centrifuged at 1,500 $\times$ *g* several times for 10 min each time to remove nuclei. The nuclei were optionally collected for solution isoelectric focusing and lysed with lysis buffer (4% SDS, 15 mM Tris-HCl (pH 7.4), and inhibitors). This supernatant was centrifuged at 20,000 $\times$ *g* for 10 min, and the pellet was washed with sucrose-Tris buffer and centrifuged again to pellet the mitochondria. This crude mitochondrial preparation was either saved for solution isoelectric focusing after treatment with lysis buffer or further purified for GELFrEE separation. The mitochondrial pellet was resuspended in several milliliters of sucrose-Tris buffer, layered over a buffered Percoll solution (50% Percoll, 10 mM Tris-MOPS, pH 7.4, 0.1 mM EGTA, and inhibitors), and centrifuged at 40,000 $\times$ *g* for 1 h. The most abundant band was carefully collected with a pipette, diluted with sucrose-Tris buffer, and centrifuged at 20,000 $\times$ *g* for 10 min. The resulting pellet was washed with sucrose-Tris buffer and then treated with lysis buffer.

Solution Isoelectric Focusing—Solution isoelectric focusing (sIEF)[1] was carried out as previously reported (3, 23). Briefly, whole cell lysate (~3 mg total protein) or preparations from subcellular fractions (~500 μg to 3 mg) were acetone precipitated and resuspended in 4 M urea, 2 M thiourea, 50 mM DTT, and 1% Bio-Lyte 3/10 carrier ampholytes (Bio-Rad, Hercules, CA) and loaded into the acrylic separation chamber of the sIEF device, which was operated at a constant 2 W for ~2 h. Eight to 10 fractions were collected from the device, and based upon the separation, as indicated through SDS-PAGE gels, fractions were pooled to obtain four sIEF fractions for subsequent fractionation via GELFrEE.

GELFrEE Separation—Protein samples (300 to 500 μg) or sIEF fractions (~50 to 300 μg) were precipitated using four volumes of acetone and suspended in 150 μl of gel loading buffer. Protein fractionation was performed using the GELFREE 8100 Fractionation System (Expedeon, Harston, Cambridgeshire, UK) using 10% or 12% T gel columns. The fractionation was visualized by silver staining of an SDS-PAGE gel with 10 μl of each GELFrEE fraction loaded onto the gel. For efficient SDS removal, the remainder of the fractions were precipitated with MeOH/$CHCl_3$/H_2O as previously described (24). The fractions were then resuspended in 20 to 30 μl of buffer A (95% H_2O, 5% acetonitrile, 0.2% formic acid).

Liquid Chromatography–Mass Spectrometry—Each sample (6.4 μl) was injected onto a 2-cm, 150-μm inner diameter PLRP-S (d_p = 5 μm, pore size = 1000 Å) trap column and washed at 3 μl/min with buffer A using a Dionex Ultimate 3000 RSLCnano system (Thermo Fisher Scientific). A 10-cm, 75-μm inner diameter PLRP-S column was used for separation. The flow rate was 300 nl/min, and a typical gradient started at 5% B (95% acetonitrile, 5% H_2O, and 0.2% formic acid) and rose to 20% B at 5 min, 65% B at 50 min, and 80% B at 58 min, where it was held for 4 min. The column was returned to 5% B over 4 min and re-equilibrated for 14 min.

Data were collected with an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) fitted with a custom nanoelectrospray source. MS^1 data were collected with either the Orbitrap analyzer, using high-resolution (60,000–120,000 resolving power at *m/z* 400) or low-resolution "express" scans (12-ms transients), or the Velos ion trap. Precursors detected by high-resolution scans were fragmented using data-dependent CID, HCD, or electron transfer dissociation. Typical MS^2 data were collected using a top-two data acquisition strategy with an isolation window of 15 *m/z* and 60,000 resolving power at *m/z* 400. Precursors detected using express or ion trap scans were fragmented using data-dependent ETD or HCD or data-independent source-induced dissociation.

Data Analysis—Intact precursor and fragment masses from LC-MS/MS files were determined using cRAWler software, which utilizes Xtract to determine monoisotopic neutral masses from high-resolution precursor and fragment ion spectra, and kDecon (25) to generate average masses from low-resolution precursor data via the deconvolution of protein charge states. The processed data were analyzed with a distributed version of ProSightPC 3.0 (Thermo Fisher Scientific) on a 168-core computing cluster. Data with high-resolution precursor masses were analyzed using a tiered, iterative, absolute mass search logic with an initial 2.3-Da precursor window and subsequent 2,000-, 20,000-, and 100,000-Da windows of MS^1 mass values. Low-resolution precursor mass data used 200-, 20,000-, and 100,000-Da windows. If a precursor mass could not be automatically determined, the entire database was searched. The iterative approach allows for a reduction in data processing time by short circuiting further, more computationally intensive searches once a statistically confident identification had been made. Fragment ions were matched using a 10-ppm mass tolerance. All searches were performed against only "reviewed" entries from the March 2013 release of UniProtKB (2013 03). Searches with smaller precursor mass windows (2.3 or 2,000 Da) were run against a highly annotated (26) PTM Warehouse (21,624,023 theoretical proteoforms), whereas searches with wider intact tolerance windows utilized a more modestly annotated warehouse containing 164,088 proteoforms. The two human databases are available for download. False discovery rates (FDRs) were determined using a *q* value estimation approach as described previously (3). Transmembrane domains were predicted using TMHMM v.2.0 (27) based upon the precise sequence identified by ProSightPC.

RESULTS

A representative total ion chromatogram for the LC-MS/MS analysis of a single GELFrEE fraction enriched for mitochondrial proteins is displayed in Fig. 1. Also shown are isotopic distributions for five mitochondrial proteins detected throughout the gradient with the expectation value (E-value) for their identification following HCD fragmentation and database searching. As illustrated, each of the five proteins was detected with <2 ppm mass error and identified with high confidence (E-value < 10^{-57}). Transmembrane protein ^{14}C (Q9P0S9) and

[1] The abbreviations used are: CID, collisionally induced dissociation; FDR, false discovery rate; GELFrEE, gel-eluted liquid fraction entrapment electrophoresis; HCD, higher-energy collisional dissociation; HMG, high mobility group; sIEF, solution isoelectric focusing.

mitochondrial SRA stem-loop-interacting RNA-binding protein (Q9GZT3) were detected with an N-terminal acetylation. ATP synthase subunit g (O75964) and the 10-kDa mitochondrial heat shock protein (P61604) each had their initial methionine residues removed and N-termini acetylated. A 22-residue transit peptide was observed to be removed from ATP synthase subunit δ (P30049). Using a top-two data-dependent acquisition method and an E-value cutoff of 10^{-4} for the search results, 153 unique proteins were identified from this single LC-MS/MS analysis. This included the identification of 67 integral membrane proteins defined by TMHMM prediction of the ProSightPC output sequence. Manual inspection of the 67 membrane protein identifications revealed that only one protein was observed as a cleaved form without a true transmembrane domain. Two proteins with five transmembrane helices were identified: translocator protein (P30536) and ATP synthase subunit a (P00846), which was observed with 26 N-terminal residues missing. Duplicate analyses of the same fraction using the same acquisition parameters but with CID fragmentation resulted in 122 and 123 identifications, respectively; 80% were the same in each run. The combination of identifications from all three LC-MS runs resulted in a combined total of 176 unique identifications from a single GELFrEE fraction.

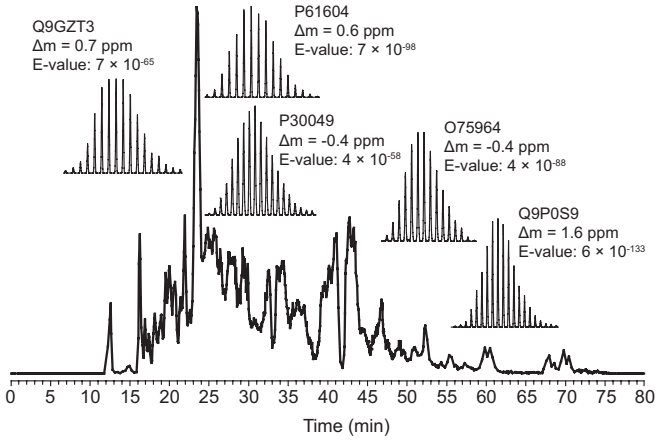

FIG. 1. **Total ion chromatogram and several identifications from a GELFrEE fraction enriched for mitochondrial proteins.** Using HCD, 153 proteins were identified in this single LC-MS/MS run.

Further comparisons were made between LC-MS runs employing CID and HCD for tandem MS of proteins in the first five fractions (<30 kDa) of the same GELFrEE separation. The combined data from these five fractions showed a ~35% increase in identifications with HCD relative to CID (supplemental Fig. S1*A*). Beyond the increase in identification number, the distributions of database retrieval scores for the most confident hits demonstrated a slightly higher confidence with HCD *versus* CID (supplemental Fig. S1*B*). The median E-value obtained with HCD was 5×10^{-23}, whereas the replicate CID sets had median E-values of 3×10^{-18} and 8×10^{-20}, respectively. In total, nearly 400 proteins were identified from five GELFrEE fractions using ~400 μg of enriched mitochondrial protein. This data set was useful in exploring the CID *versus* HCD landscape for collisional fragmentation, and it was also used to rigorously compare E-values with the scoring method via the target-decoy approach (see Table I and "Discussion"). As the *q* value is calculated based on the distribution of *p* scores, comparing a cutoff with an E-value cutoff requires knowledge of the size of the database. Two databases were utilized in the search strategies used here. Because the decoy searches were performed using the same search logic and the same databases, the database size was already considered in the *q* value approach to multiple hypothesis testing. As revealed in Table I (center row), a 1% FDR threshold of this data set corresponds to Bonferroni-corrected E-value cutoffs of 7.7×10^{-4} and 0.10 for the simple and complex human databases, respectively. In subsequent calculations, the *q* value approach was used to determine the protein-level FDR and identification count.

High-throughput Protein Identification—To extend the reach of top-down proteomics to cover a greater percentage of the human proteome, fractions highly enriched for mitochondria along with other cellular preparations (optionally prefractionated using sIEF) were analyzed through several rounds of GELFrEE-LC-MS/MS. For this study, 423 data files collected using a high-resolution precursor mass, typically <30 kDa, were analyzed. This resulted in the identification of 1,598 unique proteins using a 5% FDR threshold (corresponding to a *p* score = 3.41×10^{-8}). With a 1% FDR cutoff (*p* score = 1.33×10^{-9}), the number of unique identifications decreased from 1,598 to 1,063. Furthermore, low-resolution precursor

TABLE I
Comparison of statistical methods for identifying intact proteins in the top-down fashion

False discovery rate (*q* value cutoff) (%)	Number of unique identifications	*p* score cutoff	E-value cutoff (simple proteoform database[a])	E-value cutoff (complex proteoform database[b])
0.10	367	1.2×10^{-10}	2.0×10^{-5}	2.6×10^{-3}
1.0	407	4.7×10^{-9}	7.7×10^{-4}	0.10
5.0	491	9.2×10^{-8}	1.5×10^{-2}	2.0

[a] Proteoforms are created in a candidate expansion approach called "shotgun annotation" (26); the simple database contains 164,088 candidate proteoforms.

[b] The complex database contains 21,624,023 candidate proteoforms (see "Experimental Procedures").

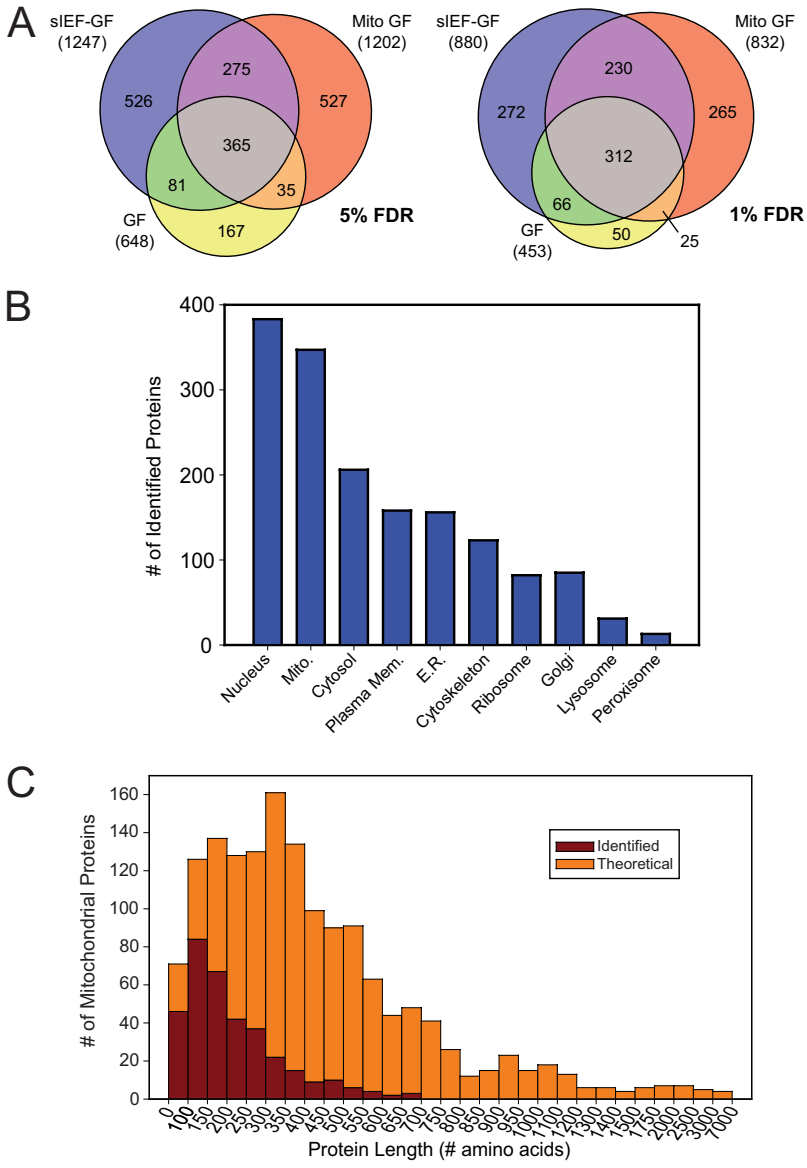

FIG. 2. **Analysis of the 1,220 total proteins identified in this study at a 1% FDR.** *A*, distribution of the identifications between GELFrEE of enriched mitochondrial proteins, IEF fractions, and other GELFrEE preparations at 5% and 1% FDRs. *B*, subcellular localization of the identified proteins revealed 347 mitochondrial proteins identified along with a significant number of proteins in most of the common organelles. *C*, distribution of all annotated human mitochondrial proteins as a function of protein length overlaid with those identified in this work.

data from 388 additional raw files were collected using either the ion trap or short transients from the Orbitrap analyzer to obtain MS1 data for proteins between 30 and 80 kDa. Using these other MS1 data types but accurate mass MS2, 905 and 578 proteins were identified using 5% (p score $= 5.05 \times 10^{-9}$) and 1% (p score $= 9.47 \times 10^{-11}$) FDR cutoffs, respectively. When the high- and low-resolution experiments were combined, a total of 1,976 unique identifications at a 5% FDR cutoff were achieved, marking this study as the largest implementation of top-down proteomics to date. At the 1% FDR cutoff, 1,220 proteins were identified. Venn diagrams highlight the number of identifications resulting from the three types of fractionation using both 5% and 1% FDR thresholds (Fig. 2*A*). All further descriptions of the proteins identified in this work reference the 1,220 identifications, which are listed in supplemental Table S1.

The subcellular localization of the identified proteins was also assessed through Gene Ontology analysis (Fig. 2*B*). Given that ribosomal proteins often have many Gene Ontology terms associated with them, we omitted them from the identified proteins annotated with nuclear and cytosolic subcellular localizations. Additionally, proteins annotated to constitute mitochondrial ribosomes were included in the mitochondrial but not the ribosomal Gene Ontology sets. Nuclear proteins, of which over 5,500 are currently listed within the "reviewed" SwissProt human database (20,265 total entries), represented the largest number of the total identified proteins. Forty-two histone proteins were identified, including macro-H2A (O75367) and five linker histone H1 proteins (H1.0, H1.2, H1.4, H1.5, and H1x), each properly identified as arising from unique genes within this multigene family. Eighty-two of the annotated ribosomal pro-

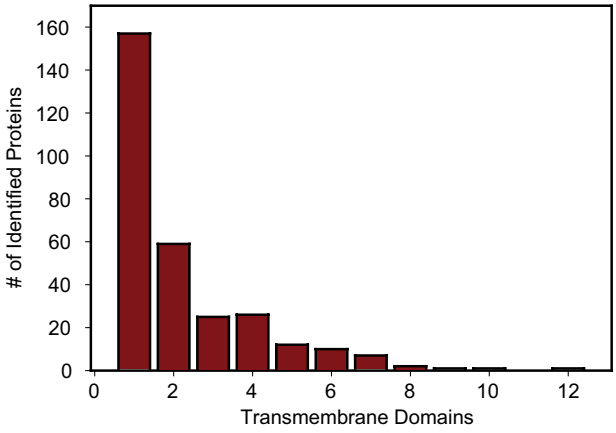

FIG. 3. **Distribution of the number of transmembrane helices predicted for the 301 integral membrane proteins identified in this study.** Proteins with between 1 and 12 transmembrane domains were identified, with 48% containing 2 or more.

teins were also identified, representing ~65% of the nonmitochondrial ribosomal proteins.

The identification of 347 mitochondrial proteins represents ~23% of all the annotated human mitochondrial proteins. However, only a fraction of the total mitochondrial proteins are expected to be expressed in any given cell type. The identified mitochondrial proteins are shown as a function of protein size in Fig. 2*C*, which illustrates high proteome coverage in the low molecular weight regime. Specifically, 47% of the annotated mitochondrial proteins under 300 residues (~33 kDa) were identified. Of all the identified mitochondrial proteins, 73 are known to be components of the five complexes of oxidative phosphorylation, representing ~78% of the total number of proteins in those complexes. All 15 subunits of ATP synthase were identified. The mitochondrial ribosome was also extensively covered, with 16 (of 32 total) and 29 (of 47 total) proteins identified from the 28S and 39S subunits, respectively.

Of the 1,220 proteins identified, 301 were found to be integral membrane proteins using TMHMM prediction of the ProSightPC output sequence. The distribution of the number of helical transmembrane domains is shown in Fig. 3. Although many of these proteins contained just one transmembrane domain, 48% were polytopic, consisting of two or more helices. A single protein with 12 transmembrane domains, cytochrome *c* oxidase subunit 1 (P00395), was identified. Based on Gene Ontology analysis, 320 proteins were determined to be"integral to membrane" and 536 "membrane," comprising integral and membrane-associated proteins.

Identification of Sequence Variants and Post-translational Modifications—Of the proteins identified in this study, many were found to have post-translational modifications or variations of primary structure. Two examples of modified proteins exhibiting multiple proteoforms within a single LC-MS run are shown in Fig. 4. Phosphorylation was detected on protein transport protein sec61 β (P60468) (Fig. 4*A*), which was localized to Ser17 using electron transfer dissociation. Additionally, the protein was observed with an unknown +58-Da difference, possibly a coding polymorphism (Ala to Glu or Gly to Asp), which was also present on the phosphorylated form. Two histidine methylations were detected on NADH dehydrogenase 1 β subcomplex subunit 3 (O43676), although the localization could not be confirmed among the expected sites, His 5, 7, and 9. The unmodified and doubly methylated proteoforms were observed at levels ~60% of those of the singly methylated form (Fig. 4*B*).

Two proteins exhibited incorrect start sites based upon UniProt annotation. Signal peptidase complex subunit 1 (Q9Y6A9) was identified with Met68 as the true site for translational initiation (Fig. 5*A*), matching with form BAG51320A.1 from the European Bioinformatics Institute. The Met27 residue of keratinocyte associated protein 2 (Q8N6L1) was found as the true start site, which was then cleaved. A proteoform with a −14-Da intact mass difference in cytochrome b-c1 complex subunit 9 (Q9UDW1) was observed, present at a level about equal with the main peak; this proteoform was attributed to a known I47V coding polymorphism with a 0.002-Da (0.2-ppm) mass error. Several proteins displayed differences in signal sequence cleavage, based on MS1 and MS2 evidence, from that currently annotated within UniProt, including the removal of an additional serine from the N terminus of NADH dehydrogenase flavoprotein 3 (P56181). Ras-related protein rap-1b (P61224) was found to have the last three terminal residues cleaved and a methyl ester on its C terminus as annotated within UniProt (Fig. 5*B*). Lipid modifications were also detected, including myristoylation found on the N-terminal glycine residues of 11 proteins including ADP-ribosylation factors 1, 3, 4, and 5 (P84077, P61204, P18085, P84085), calcineurin B homologous protein 1 (Q99653), and MARCKS-related protein (P49006). Myristoylation of plasminogen receptor (Q9HBL7) was found manually from a +210-Da intact mass shift (+210.2 Da, theoretical) and was confirmed further with MS2 data. This modification is not currently annotated and is believed to be novel to this study. An unannotated palmitoylation was detected on Golgi vesicular membrane-trafficking protein p18 (O15155). As evident from the mass shift and protein fragmentation pattern, the probable localization of this new palmitoylation is cysteine 98 (data not shown). A known farnesylation, a sesquiterpene modification, was detected on the C-terminal cysteine of GTP binding protein Rheb (Q15382). Geranylgeranylation, a 20-carbon diterpene modification, was detected on both of two adjacent cysteine residues of both Ras-related protein Rab-11a (P62491) and Rab-11b (Q15907), which also both featured a cleaved initial methionine and N-terminal acetylation (Fig. 5*C*). Guanine nucleotide-binding protein subunit γ-12 (Q9UBI6) exhibited cleavage of its last three residues, with geranylgeranylation of the C-terminal cysteine as well as C-terminal methylation. Similarly, Ras-related protein Rab-14 (P61106) exhibited C-terminal methylation and geranylgeranylation of either the C-

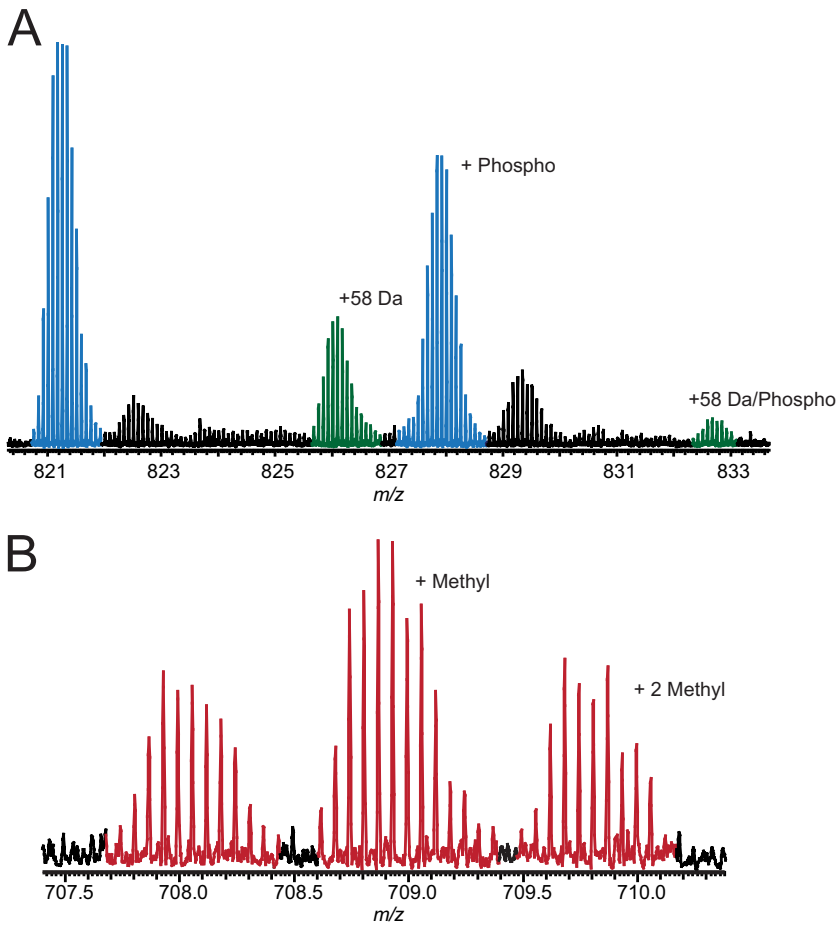

FIG. 4. **Detection of multiply modified proteins.** *A*, phosphorylation of protein transport protein sec61 β (P60468), which was also observed with a +58 Da form, also phopsphorylated. *B*, NADH dehydrogenase 1 β subcomplex subunit 3 was detected with two histidine methylations along with the unmodified form, with the singly methylated form being the most abundant.

terminal cysteine (Cys 215) or Cys 213. Phosphorylation was found on more than 20 proteins, including anaphase-promoting complex subunit CDC26 (Q8NHZ8) and prostaglandin E synthase 3 (Q15185), which was found to be triply phosphorylated. Lysine trimethylation was detected on calmodulin and ATP synthase subunit c, which remains unannotated in UniProt but was previously reported elsewhere (5, 28). Monomethylation of lysine 5 of 60S ribosomal protein L29 (P47914) was also identified. Dimethylation of arginine 156 and 158 was observed on 40S ribosomal protein S10 (P46783). All of the forms discussed above are based on fragmentation evidence, often with intact mass differences further supporting the characterization.

Proteoform Changes in Cellular Senescence—With increased coverage of the human proteome, it becomes possible to screen for proteoform-level changes related to chemically induced cellular senescence. A Venn diagram comparing the identifications of proteins extracted from treated (276 LC-MS runs) and control cells (535 runs) is shown in Fig. 6*A*. A variety of proteins were observed exclusively in treated cells, including macro-H2A (O75367), known to be enriched in senescence-associated heterochromatin foci (29), and histone H1x (Q92522). Additionally, several proteins involved in the regulation of the cell cycle, including cell cycle progression protein 1 (Q9ULG6) and G2/mitotic-specific cyclin-B1 (P14635), and in the DNA damage response, including ATP-dependent DNA helicase Q1 (P46063) and centrosomal protein of 63 kDa (Q96MT8), were found exclusively in the treated cells. Ragulator complex protein LAMTOR5 (O43504), known to suppress the mitochondrial pathway of apoptosis (30), was also identified only in the treated cell population. Phosphatidylinositol N-acetylglucoasminyltransferase subunit P (P57054) and protein preY (Q96I23), required for the production of glycosylphosphatidylinositol (31), were found and may play a role in senescence (32).

Additionally, hyperphosphorylated forms of HMGA2 were found (Fig. 6*C*). Because of the lower abundance of HMGA2 relative to HMGA1 (see Fig. 6*B*), whose changes during senescence were described previously (3), the analysis of HMGA2 was shifted from discovery to targeted mode to improve the quality of MS1 and MS2 data. In the control population, a distribution from zero to three phosphorylations was observed with the major proteoform exhibiting a single phosphorylation (Fig. 6*C*, upper panel). However, upon treatment the proteoform distribution shifted, with up to five phosphorylations observed (Fig. 6*C*, lower panel). Only 3 sites of phosphorylation are currently annotated within UniProt, whereas 12 sites are listed on PhosphoSitePlus. All proteoforms observed for this

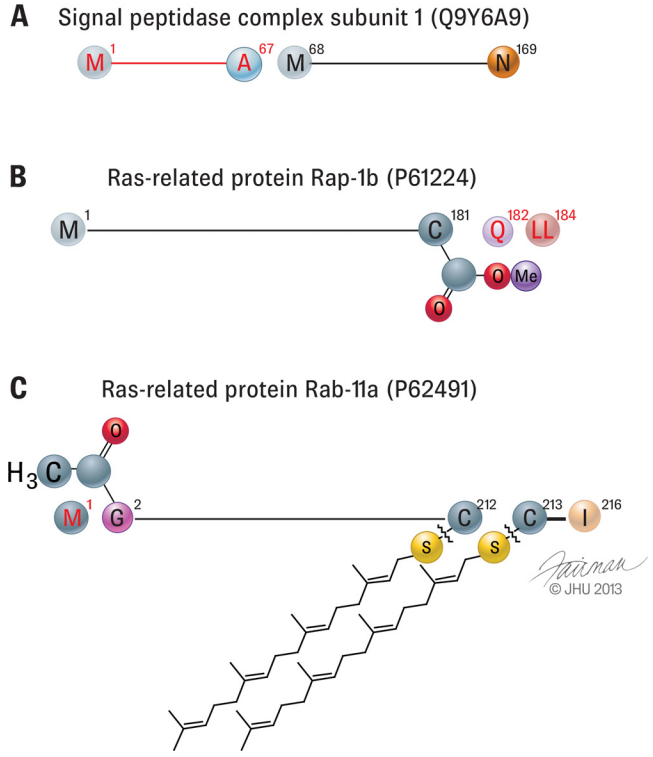

FIG. 5. **Identification of post-translational modifications and sequence variants.** *A*, signal peptidase complex subunit 1 (Q9Y6A9) was found as a shortened form, with the predicted Met68 as the initial methionine. *B*, Ras-related protein Rap-1b (P61224) showed cleavage of its three C-terminal residues and methylation of the C terminus. *C*, Ras-related protein Rab-11a (P62491) exhibited geranylgeranylation on two adjacent cysteine residues as well as cleavage of the initial methionine and acetylation of the N terminus.

protein had the initial methionine cleaved and N-terminal acetylation.

Investigation began into the phosphorylation hierarchy of HMGA2 proteoforms. Isolating the singly and doubly phosphorylated forms shown in Fig. 6*C* (upper panel) allowed us to localize the first phosphorylation to S105 (supplemental Fig. S2, top). The second appeared distributed among multiple sites but could not be definitively localized to the two others annotated in SwissProt (T40 and S44). Low phosphorylation levels necessitated the characterization of phosphorylated forms from senescent cells (Fig. 6*C*, lower panel). In the triply phosphorylated form, all three phosphorylations were present on the highly acidic C-terminal tail; drawing on HMGA1 as an example, these sites are likely S105, S102, and S101. The fourth phosphorylation site was present on S44, and the fifth appears to occur at T40 (supplemental Fig. S2).

DISCUSSION

In this top-down work, featuring the highest human proteome coverage to date, we identified 1,220 proteins with the assistance of several different fractionation procedures. The incorporation of the Orbitrap Elite offered superior resolution and speed relative to the instruments used previously. Additionally, the use of HCD and electron transfer dissociation improved fragmentation relative to traditional CID. Using mitochondrial enriched samples along with sIEF-GELFrEE fractionation of whole cell lysate and subcellular fractions, we showed that substantial coverage of the low molecular weight human proteome is possible without proteolytic digestion. The number of identifications presented here is greater than that from the most expansive top-down report thus far and is more selective, with a 1% FDR cutoff rather than a 5% cutoff (3). Comparing the identifications, at a 5% FDR, this work shows an 89% increase in the scale of protein identifications. Expanding proteome coverage led to the identification of 301 integral membrane proteins, the most reported to date, without membrane enrichment or modified solubilization conditions. Additionally, many of the identified proteins showed post-translational modifications, including a variety of lipid modifications, some previously unknown. Specifically regarding the mitochondria, 347 proteins annotated as mitochondrial proteins were identified with ~50% of the predicted mitochondrial proteome under 30 kDa. For comparison, the previous largest top-down study identified 186 mitochondrial proteins (3). The furthest reaching bottom-up study of a mammalian proteome resulted in a highly refined mitochondrial data set consisting of 1,098 proteins from 14 different mouse tissues. The bottom-up analysis of single tissues has resulted in between ~550 and 800 protein identifications. Although top-down proteomics will require further technical development, especially at higher masses, in order to achieve the identification levels obtainable from a peptide-based approach, the identifications achieved here show that comparable coverage between the two approaches is conceivable. Highly purified mitochondria from mouse tissues may serve as tractable samples for comparison of the two proteomics approaches.

Comparison of data collected in technical duplicate utilizing CID showed ~80% reproducibility of identifications, with a single analysis utilizing HCD capturing ~95% of the identifications from both CID replicates, highlighting the reproducibility of the analysis. Overall, the utility of HCD intact protein dissociation was demonstrated with a 35% increase in identifications while improving the average confidence of the identifications, further highlighting the benefits of Orbitrap instruments for top-down proteomics. The same data set was used to compare two statistical methods for scoring database search results: (1) E-values based on a Bonferroni-corrected Poisson model (33), and (2) a target-decoy method that produces instantaneous FDRs, or *q* values (34). This allowed us to survey the identification landscape around the threshold used to count identifications (Table I). Given that Bonferroni correction is the most conservative approach to setting such thresholds for counting protein identifications, the 1×10^{-4} cutoff used in past work is highly conservative, especially for a highly annotated human database. The *q* value approach currently presents the best way to temper

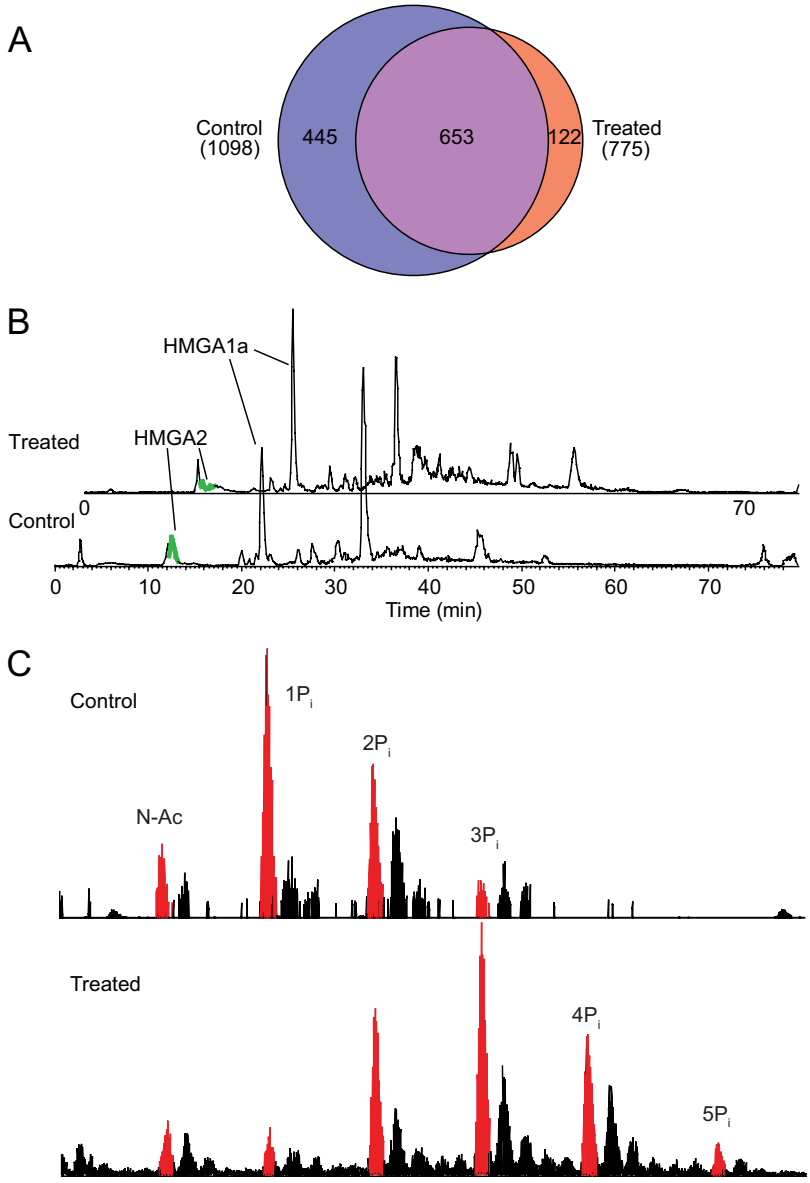

FIG. 6. **Comparison of untreated and senescent H1299 cells at the protein identification and proteoform levels.** *A*, Venn diagram displaying the protein identifications found in the two cell states. *B*, detection of HMGA2 in H1299 control and senescent cells. HMGA2 was detected through GELFrEE-LC-MS and showed significantly lower abundance than HMGA1. *C*, the phosphorylated proteoforms of HMGA2 are dynamic during senescence, with forms containing up to five phosphorylations visible (lower panel). The smaller peaks to the right of each major peak represent oxidation (+16), not methylation (+14).

claims of proteome coverage by more accurately estimating FDRs. Adoption of this approach by the top-down proteomics community could allow for more accurate comparisons between various labs.

The study of chemically induced senescence resulted in the identification of several proteins solely in the treated data set; this does not provide quantitative evidence of a biological difference, but it does lead to several targets that will be further investigated. Clearly there is a need for improved bioinformatic tools to score and mine the data-rich repositories of proteoforms that can be generated now. Additionally, it will be essential for the top-down community to develop a robust platform for intact protein quantitation that can quantify the relative abundance of proteoforms between treatments, rather than solely quantifying relative levels of modified proteoforms. These future quantitative proteome studies will highlight proteoform-level changes between cellular or phenotypic states. With true quantitation, through labeling or label-free methods, coupled with proteoform-resolved information, top-down mass spectrometry will provide more insights into the protein-level regulation and dynamics of complex biological systems.

Acknowledgments—A.D.C. acknowledges the ACS Division of Analytical Chemistry and the Society for Analytical Chemists of Pittsburgh (SACP) for their support.

* This work was supported by NIGMS, National Institutes of Health, under award number R01 GM067193 (N.L.K.). Additional support was provided by the UIUC Center for Neuroproteomics on Cell to Cell Signaling (P30 DA018310), the Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust, and the Robert H. Lurie Comprehensive Cancer Center.

[S] This article contains supplemental material.

§ To whom correspondence should be addressed: Prof. Neil

Kelleher, Tel.: 1-847-467-4362, Fax: 1-847-467-3276; E-mail: n-kelleher@northwestern.edu.

REFERENCES

1. Kelleher, N. L. (2004) Top-down proteomics. *Anal. Chem.* **76,** 196A-203A.
2. Chait, B. T. (2006) Mass spectrometry: bottom-up or top-down? *Science* **314,** 65–66
3. 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. M., Early, B. P., Siuti, N., LeDuc, R. D., Compton, P. D., Thomas, P. M., and Kelleher, N. L. (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480,** 254–258
4. Kellie, J. F., Catherman, A. D., Durbin, K. R., Tran, J. C., Tipton, J. D., Norris, J. L., Witkowski, C. E., Thomas, P. M., and Kelleher, N. L. (2012) Robust analysis of the yeast proteome under 50 kDa by molecular-mass-based fractionation and top-down mass spectrometry. *Anal. Chem.* **84,** 209–215
5. Catherman, A. D., Li, M., Tran, J. C., Durbin, K. R., Compton, P. D., Early, B. P., Thomas, P. M., and Kelleher, N. L. (2013) Top down proteomics of human membrane proteins from enriched mitochondrial fractions. *Anal. Chem.* **85,** 1880–1888
6. Ahlf, D. R., Compton, P. D., Tran, J. C., Early, B. P., Thomas, P. M., and Kelleher, N. L. (2012) Evaluation of the compact high-field Orbitrap for top-down proteomics of human cells. *J. Proteome Res.* **11,** 4308–4314
7. Ansong, C., Wu, S., Meng, D., Liu, X., Brewer, H. M., Deatherage Kaiser, B. L., Nakayasu, E. S., Cort, J. R., Pevzner, P., Smith, R. D., Heffron, F., Adkins, J. N., and Paša-Tolić, L. (2013) 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.*
8. Smith, L. M., and Kelleher, N. L. (2013) Proteoform: a single term describing protein complexity. *Nat. Methods* **10,** 186–187
9. Molnar, K. S., Karabacak, N. M., Johnson, J. L., Wang, Q., Tiwari, A., Hayward, L. J., Coales, S. J., Hamuro, Y., and Agar, J. N. (2009) A common property of amyotrophic lateral sclerosis-associated variants: destabilization of the copper/zinc superoxide dismutase electrostatic loop. *J. Biol. Chem.* **284,** 30965–30973
10. Dong, X., Sumandea, C. A., Chen, Y.-C., Garcia-Cazarin, M. L., Zhang, J., Balke, C. W., Sumandea, M. P., and Ge, Y. (2011) Augmented phosphorylation of cardiac troponin I in hypertensive heart failure. *J. Biol. Chem.*
11. 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., and Duménil, G. (2011) Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination. *Science* **331,** 778–782
12. Carroll, J., Fearnley, I. M., and Walker, J. E. (2006) Definition of the mitochondrial proteome by measurement of molecular masses of membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 16170–16175
13. Carroll, J., Altman, M. C., Fearnley, I. M., and Walker, J. E. (2007) Identification of membrane proteins by tandem mass spectrometry of protein ions. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 14330–14335
14. Ryan, C. M., Souda, P., Bassilian, S., Ujwal, R., Zhang, J., Abramson, J., Ping, P. P., Durazo, A., Bowie, J. U., Hasan, S. S., Baniulis, D., Cramer, W. A., Faull, K. F., and Whitelegge, J. P. (2010) Post-translational modifications of integral membrane proteins resolved by top-down Fourier transform mass spectrometry with collisionally activated dissociation. *Mol. Cell. Proteomics* **9,** 791–803
15. Thangaraj, B., Ryan, C. M., Souda, P., Krause, K., Faull, K. F., Weber, A. P. M., Fromme, P., and Whitelegge, J. P. (2010) Data-directed top-down Fourier-transform mass spectrometry of a large integral membrane protein complex: photosystem II from Galdieria sulphuraria. *Proteomics* **10,** 3644–3656
16. Whitelegge, J., Halgand, F., Souda, P., and Zabrouskov, V. (2006) Top-down mass spectrometry of integral membrane proteins. *Expert Rev. Proteomics* **3,** 585–596
17. Taylor, S. W., Warnock, D. E., Glenn, G. M., Zhang, B., Fahy, E., Gaucher, S. P., Capaldi, R. A., Gibson, B. W., and Ghosh, S. S. (2002) An alternative strategy to determine the mitochondrial proteome using sucrose gradient fractionation and 1D PAGE on highly purified human heart mitochondria. *J. Proteome Res.* **1,** 451–458
18. Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H. E., Schonfisch, B., Perschil, I., Chacinska, A., Guiard, B., Rehling, P., Pfanner, N., and Meisinger, C. (2003) The proteome of Saccharomyces cerevisiae mitochondria. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 13207–13212
19. Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., Patterson, N., Lander, E. S., and Mann, M. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115,** 629–640
20. Forner, F., Foster, L. J., Campanaro, S., Valle, G., and Mann, M. (2006) Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol. Cell. Proteomics* **5,** 608–619
21. Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P., Ignatchenko, A., Scott, M. S., Gramolini, A. O., Morris, Q., Hallett, M. T., Rossant, J., Hughes, T. R., Frey, B., and Emili, A. (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125,** 173–186
22. Debacq-Chainiaux, F., Erusalimsky, J. D., Campisi, J., and Toussaint, O. (2009) Protocols to detect senescence-associated beta-galactosidase (SA-βgal) activity, a biomarker of senescent cells in culture and in vivo. *Nat. Protoc.* **4,** 1798–1806
23. Tran, J. C., and Doucette, A. A. (2008) Rapid and effective focusing in a carrier ampholyte solution isoelectric focusing system: a proteome prefractionation tool. *J. Proteome Res.* **7,** 1761–1766
24. Wessel, D., and Flugge, U. I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138,** 141–143
25. Durbin, K. R., Tran, J. C., Zamdborg, L., Sweet, S. M. M., Catherman, A. D., Lee, J. E., Li, M. X., Kellie, J. F., and Kelleher, N. L. (2010) Intact mass detection, interpretation, and visualization to automate Top-Down proteomics on a large scale. *Proteomics* **10,** 3589–3597
26. Pesavento, J. J., Kim, Y.-B., Taylor, G. K., and Kelleher, N. L. (2004) Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry. *J. Am. Chem. Soc.* **126,** 3386–3387
27. Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182
28. Chen, R. M., Fearnley, I. M., Palmer, D. N., and Walker, J. E. (2004) Lysine 43 is trimethylated in subunit c from bovine mitochondrial ATP synthase and in storage bodies associated with Batten disease. *J. Biol. Chem.* **279,** 21883–21887
29. Zhang, R., Poustovoitov, M. V., Ye, X., Santos, H. A., Chen, W., Daganzo, S. M., Erzberger, J. P., Serebriiskii, I. G., Canutescu, A. A., Dunbrack, R. L., Pehrson, J. R., Berger, J. M., Kaufman, P. D., and Adams, P. D. (2005) Formation of macroH2A-containing senescence-associated heterochromatin foci and senescence driven by ASF1a and HIRA. *Dev. Cell* **8,** 19–30
30. Marusawa, H., Matsuzawa, S.-I., Welsh, K., Zou, H., Armstrong, R., Tamm, I., and Reed, J. C. (2003) HBXIP functions as a cofactor of survivin in apoptosis suppression. *EMBO J.* **22,** 2729–2740
31. Murakami, Y., Siripanyaphinyo, U., Hong, Y., Tashima, Y., Maeda, Y., and Kinoshita, T. (2005) The initial enzyme for glycosylphosphatidylinositol biosynthesis requires PIG-Y, a seventh component. *Mol. Biol. Cell* **16,** 5236–5246
32. Kooyman, D. L., Byrne, G. W., and Logan, J. S. (1998) Glycosyl phosphatidylinositol anchor. *Exp. Nephrol.* **6,** 148–151
33. Meng, F. Y., Cargile, B. J., Miller, L. M., Forbes, A. J., Johnson, J. R., and Kelleher, N. L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **19,** 952–957
34. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 9440–9445