Quantitative Top-Down Proteomics of SILAC Labeled Human Embryonic Stem Cells

Timothy S. Collier,[a] Prasenjit Sarkar,[b] Balaji Rao,[b] and David C. Muddiman[a]

[a] W. M. Keck FT-ICR Mass Spectrometry Laboratory, Department of Chemistry, North Carolina State University, Raleigh, North Carolina, USA
[b] Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina, USA

Human embryonic stem cells (hESCs) are self-renewing pluripotent cells with relevance to treatment of numerous medical conditions. However, a global understanding of the role of the hESC proteome in maintaining pluripotency or triggering differentiation is still largely lacking. The emergence of top-down proteomics has facilitated the identification and characterization of intact protein forms that are not readily apparent in bottom-up studies. Combined with metabolic labeling techniques such as stable isotope labeling by amino acids in cell culture (SILAC), quantitative comparison of intact protein expression under differing experimental conditions is possible. Herein, quantitative top-down proteomics of hESCs is demonstrated using the SILAC method and nano-flow reverse phase chromatography directly coupled to a linear-ion-trap Fourier transform ion cyclotron resonance mass spectrometer (nLC-LTQ-FT-ICR-MS). In this study, which to the best of our knowledge represents the first top-down analysis of hESCs, we have confidently identified 11 proteins by accurate intact mass, MS/MS, and amino acid counting facilitated by SILAC labeling. Although quantification is challenging due to the incorporation of multiple labeled amino acids (i.e., lysine and arginine) and arginine to proline conversion, we are able to quantitatively account for these phenomena using a mathematical model. (J Am Soc Mass Spectrom 2010, 21, 879–889) © 2010 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

Human embryonic stem cells (hESCs) are pluripotent cells derived from the inner cell mass of blastocyst stage embryos that possess the ability to self-renew or differentiate into almost any human cell type. This makes hESCs a potentially valuable tool for regenerative medicine, where tissues damaged by disease or injury can be restored via hESC-derived cells. The ability of hESCs to be maintained in cell culture allows the opportunity for investigation into how these cells maintain pluripotency and how they progress through differentiation to various cell types [1]. Exploring the proteome of hESCs and how it changes qualitatively and quantitatively is an integral part of understanding the overall biology of this cell type, and will yield invaluable information toward the development of new medical treatments.

Top-down proteomics has emerged as a powerful technique for protein analysis and is a growing area of research in the mass spectrometry community. The typical experimental strategy for top-down proteomic analysis includes the front-end separation of intact proteins, their detection and fragmentation in a mass spectrometer, and subsequent identification based on sequence information obtained from their resulting MS and MS/MS spectra [2, 3]. Intact proteins present a wide range of physical and chemical properties that challenge front-end separation techniques [4–6], mass analyzer duty cycle [7], and protein identification algorithms. These challenges hinder high throughput analyses and, thus, the top-down approach has lagged the more commonly employed bottom-up approach. However, the emergence and development of new separations technologies such as GELFREE separation [8–11], high resolving power hybrid mass analyzers such as the LTQ-FT-ICR [12], and LTQ-Orbitrap [13, 14], and bioinformatics software such as ProSightPC [15] have greatly improved the capability for addressing the challenges posed by top-down proteomics. Recently, the top-down approach has been utilized by Kelleher and coworkers to examine multiple modifications of human histones, including methylation and acetylation [16]. The Kelleher group also has applied GELFREE technology to address front end separation challenges in the top-down analysis of the low mass proteome of HeLa cell extracts [17]. Top-down proteomics has also been used to identify single amino acid polymorphisms and characterize post-translational proteolysis and phosphorylation of affinity purified rat cardiac troponin [18].

Address reprint requests to Dr. D. C. Muddiman, Keck FT-ICR Mass Spectrometry Laboratory, Department of Chemistry, North Carolina State University, Raleigh, NC 27695, USA. E-mail: david_muddiman@ncsu.edu

Integrated top-down and bottom-up strategies have allowed investigators to take advantage of the top-down approach's ability to identify and characterize multiple post-translational modifications (PTMs) and identify genetic modifications while utilizing bottom-up proteomics to achieve higher throughput [19].

Quantitative top-down proteomics most often utilizes stable isotope labeling to create an internal standard from which reliable quantitative data may be obtained. Gordon et al. demonstrated with bovine and equine cytochrome *c* that even small differences-in this case, three amino acids between a protein and a structural homologue from a different species are enough to produce different charge state distributions that hinder quantification unless the signal from all charge states is taken into account [20]. Deuterated leucine has been incorporated into model cell cultures, including *E. coli* and *S. cerevisiae* [21, 22], which facilitated identification of proteins using accurate intact mass and amino acid counting; however, quantification with deuterated labels using reversed-phase chromatography is problematic because elution times decrease for the deuterated analytes [23, 24]. $^{15}N/^{14}N$ metabolic labeling has been successfully utilized the yeast proteome and, moreover, was employed on a chromatographic time scale to identify 22 intact proteins from *S. cerevisiae* [25]. Stable isotope labeling by amino acids in cell culture (SILAC) was introduced by Mann and coworkers as another means of quantifying peptides and proteins, whereby stable isotope labeled amino acids are supplemented to cull culture growth media to produce co-eluting labeled and unlabeled protein or peptide analytes. This technique is widely implemented in bottom-up experiments using labeled arginine and lysine residues in conjunction with trypsin proteolysis [26, 27].

Quantification of intact proteins using SILAC has been computationally modeled and demonstrated in affinity purified proteins expressed with an *E. coli* vector and 99% stable isotope incorporation of a single label was achieved [28]. We have previously described the implementation of SILAC on a chromatographic time scale for protein identification and quantification using $^{13}C_6$-arginine supplemented with the growth media of the filamentous fungus *Aspergillus flavus*, and also demonstrated that incomplete label incorporation in a prototrophic organism results in broad protein isotopic distributions that hinders the use of SILAC labeled intact proteins for relative quantification or assistance in identification. A mathematical relationship was also described, which characterizes the relationship of label incorporation with the observed experimental protein signal [29].

Mass spectrometry based proteomics is becoming increasingly employed for the analysis of various stem cell populations at increasing depth and detail [30]. Recently, mass spectrometry, electron capture, and electron-transfer dissociation (ECD and ETD, respectively) were used to characterize the phosphoproteome of hESCs [31, 32] and label free quantification has been carried out to characterize 74 different PTM combinations of histone H4 in hESCs [33]. The iTRAQ method has been used to track the change in mouse embryonic stem cell protein expression during differentiation [34]. Bottom-up quantification using SILAC has been utilized to identify and quantify over 5000 mouse embryonic stem cell proteins [35], and has been recently demonstrated with hESCs to seek out differentially expressed cell surface markers between pluripotent and differentiating cell populations [36].

This study represents our initial efforts to establish top-down analysis of SILAC labeled human embryonic stem cells, in which we utilize top-down proteomics to analyze human embryonic stem cells that have been dual labeled with $^{13}C_6$-arginine (Arg6) and $^{13}C_6, ^{15}N_2$-lysine (Lys8) with approximately greater than 98% label incorporation as verified by bottom-up proteomics. We detected 62 intact protein precursors from which we were able to unambiguously identify 11 based on accurate intact mass, MS/MS fragmentation data, and amino acid counting. Furthermore, we characterize the effect of multiple stable isotope labels, minute incorporation of unlabeled amino acids, and arginine-proline conversion on the labeled protein's isotopic distribution.

Experimental

Culture and SILAC Labeling of hESCs

H9 cells were maintained on mouse embryonic fibroblast (MEF) conditioned medium (CM) according to a protocol adapted from Xu et al. [37]. Briefly, MEF cells were inactivated by mitomycin *c* (Sigma Aldrich, St. Louis, MO, USA) treatment and plated at 55,000/cm². HESC medium consisting of DMEM/F12 (Invitrogen, Carlsbad, CA, USA), 20% knockout serum replacement (Invitrogen), 1 mM L-glutamine (Invitrogen), 0.1 mM β-mercaptoethanol, 1% nonessential amino acids (Invitrogen), and 4 ng/mL basic fibroblast growth factor (bFGF; Sigma Aldrich) was conditioned for 24 h on MEF cells. CM was collected for 7 d, pooled, and supplemented with 4 ng/mL bFGF (Sigma Aldrich).

CM for SILAC labeling of hESCs was prepared using protocols previously described [38]. Briefly, SILAC medium consisting of DMEM/F12 without L-lysine and L-arginine (Pierce, Rockford, IL, USA), $^{13}C_6$ L-arginine (Pierce), $^{13}C_6$ $^{15}N_2$ L-Lysine (Pierce), 20% knockout serum replacement (Invitrogen), 1 mM L-Glutamine (Invitrogen), 0.1 mM β-mercaptoethanol, 1% nonessential amino acids (Invitrogen), and 4 ng/mL basic fibroblast growth factor (bFGF, Sigma Aldrich) was conditioned for 24 h on MEF cells. CM was collected for 7 d, pooled, and supplemented with 4 ng/mL bFGF (Sigma Aldrich). It is important to note that the knockout serum replacement does not contain lysine or arginine [39].

Protein Isolation and Sample Preparation

Confluent H9 cells were lysed with 8M Urea (Sigma Aldrich) and 50 mM ammonium bicarbonate (Fisher Scientific, Hampton, NH, USA) and frozen at −80 °C. Before analysis by LC-MS, cell lysates were concentrated and desalted on 5 μm, C_8 stationary phase beads (Michrom Bioresources, Auburn, CA, USA) according to a protocol adapted from Winston and Fitzgerald [40] in which 5 mg of stationary phase beads in 400 μL of methanol was added to ~75 μL of intact protein SILAC mixture in a 1.5 mL microcentrifuge tube and vortexed for 5 min, followed by 3 min centrifugation at 10,000 × g. After rinsing the sides of the tube twice with 20 μL of methanol, the sample was centrifuged again for 3 min at 10,000 × g. After discarding the supernatant, the protein-stationary phase pellet was washed with 1 mL of 0.1% trifluoroacetic acid (Sigma Aldrich), vortexed for 1 min, and then centrifuged for 3 min at 10,000 × g. The steps were repeated two more times before reconstituting sample in 100 mM ammonium bicarbonate pH 7.8 and filtering through a 50 kDa Microcon molecular weight cutoff filter (Millipore, Billerica, MA, USA) to remove stationary phase beads from solution.

Bottom-up sample preparation. Fifty μg of Arg6, Lys8-labeled protein from human embryonic stem cells was combined with 2× Laemmli buffer containing β-mercaptoethanol, boiled 5 min, and cooled on ice. The sample was loaded and ran on a Criterion 10%–20% Tris-HCl gel (BioRad, Hercules, CA, USA) at 200 V and visualized with BioSafe Coomassie stain. The gel lane was divided into eight equal fractions, each of which was reduced with DTT, alkylated with iodoacetamide, and digested with trypsin according to a protocol adapted from Shevchenko et al. [41] First, excised gel bands were incubated for 30 min at room temperature in a 1:1 mixture of acetonitrile:100 mM ammonium bicarbonate pH 7.8. After removing the solution to waste, the gel pieces were dehydrated with 100% acetonitrile. Gel pieces were reduced for 30 min at 56 °C in 10 mM dithiothreitol (Sigma Aldrich) in 100 mM ammonium bicarbonate pH 7.8, followed by dehydration with 100% acetonitrile. Free thiols from the reduced cys–cys disulfide bonds were alkylated with 55 mM iodoacetamide (Sigma Aldrich) in 100 mM ammonium bicarbonate and incubated in the dark at room temperature. After dehydration with acetonitrile, samples were digested for 18 h at 37 °C with 20 ng/μL proteomics grade trypsin (Sigma Aldrich) in 36 mM ammonium bicarbonate, 0.1 mM HCl, and 9% acetonitrile. The digestion was halted with the addition of a 1:2 mixture of 5% formic acid:acetonitrile followed by incubation for 15 min at 37 °C, after which the solution containing the tryptic peptides was removed and saved. The gel pieces were dehydrated with acetonitrile, and the resulting solution was removed and pulled with the other samples, which were reduced to dryness in a SpeedVac

(ThermoFisher, San Jose, CA, USA) and frozen at −80 °C until needed.

Top-down LC-MS/MS. All LC solvents were purchased from Burdick and Jackson (Muskegon, MI, USA). Reversed-phase nano-flow liquid chromatography was performed on a nanoLC-2D system (Eksigent, Dublin, CA, USA) utilizing a vented column configuration [42]. Ten μL samples containing ~1 μg of total protein were injected using the autosampler, desalted on a 75 μm × 50 mm IntegraFrit (New Objective, Woburn, MA, USA) trap packed in-house with mRP-C18 5 μm stationary phase and eluted onto a PicoFrit capillary column with a 15 μm emitter tip (New Objective) also packed in-house with mRP-C18 stationary phase (Agilent, Palo Alto, CA, USA). The final packed volume of the column had dimensions 75 μm i.d. × 200 mm and was enclosed in a modified Hot-Pocket column heater (ThermoFisher) operating at 80 °C. Mobile phases A and B consisted of 98% water, 2% acetonitrile, and 2% water, 98% acetonitrile, respectively, with both containing 0.2% (vol/vol) formic acid as an ion-pairing reagent. Samples were injected and desalted on the trap at 1.5 μL/min with 2% mobile phase B for a metered injection of 30 μL. After desalting, a 10-port valve (VICI, Houston, TX, USA) switched the trap in line with a nano-flow gradient at 300 nL/min, at which time data collection also commenced. The gradient held at 2% mobile phase B for 10 min, ramped to 25% over 5 min, then to 65% B over the following 70 min. The gradient was then ramped to 95% B in 5 min and held for 10 min before re-equilibrating at 2% B for 10 min.

Mass spectrometric analyses were performed on a 7T LTQ-FT Ultra (ThermoFisher) with a pulse sequence consisting of four events, providing for a broadband acquisition in profile mode followed by three data-dependent MS/MS experiments. All mass measurements were performed inside the ICR cell. All events used 1 microscan to determine ionization time to reach the target AGC limit of 1×10^6. The resolving power of all four events was set at $100,000_{FWHM}$ at m/z = 400. MS/MS settings used an isolation width of 7 m/z and a normalized collision energy of 18% for 30 ms. CID was performed on the three most abundant m/z values from the precursor ion scan in the LTQ followed by product ion detection in the ICR cell. Precursor ions with 1+, 2+, and 3+ charges were excluded from MS/MS. Precursors isolated for MS/MS were dynamically excluded for 45 s after three repeat analyses within a 30 s window to reduce redundant analysis of abundant precursor ions that may dominate the mass spectra for long periods of time.

Top-down data analysis. RAW LC-MS data files were imported into ProSightPC 2.0 (ThermoFisher) using a THRASH [5] algorithm to determine the monoisotopic peak of analyte signals with a S/N ratio of 3:1 or greater. Intact precursor and fragment ion masses were searched against the IPI human database (ver. 3.63)

using a search logic tree that submitted all data to an absolute mass search, followed by a biomarker search for all data that did not yield an ID with E value less than $1E^{-4}$ in the absolute mass search. Intact protein precursor and fragment ion data were searched initially with ion mass tolerances set at ±10 ppm. An additional search was performed with fragment tolerances of ±1 Da to accommodate sample handling modifications to the mass (e.g., deamidation) or error in monoisotopic peak picking by the THRASH algorithm. Identified protein sequences reported from the searches were confirmed using arginine and lysine counting from the precursor ion spectra. Analysis of SILAC labeled protein distributions and theoretical distributions with varying arginine and lysine labeling efficiencies in addition to different amounts of proline conversion was carried out using ICR-2LS (PNNL, Richland, WA, USA) and Microsoft Excel (Redmond, WA, USA).

Bottom-up LC-MS/MS. In-gel digested peptide samples were separated using the same LC system, column configuration and mobile phases described vide supra. The stationary phase employed was Magic C18AQ (5 μm particle, 200 Å pore size) (Microm BioResources, Auburn, CA, USA). Ten μL of sample was injected and desalted at 1.5 μL/min on the trap before switching in line with the gradient at a flow rate of 300 nL/min. The gradient held at 2% B for 5 min before ramping to 10% B over 2 min, followed by an increase to 50% B over the next 60 min. The gradient then ramped to 95% B in 3 min, held for 5 min, then re-equilibrated at 2% B for 10 min.

Peptides were detected on the LTQ-FT Ultra with a pulse sequence containing four scan events. A broadband scan was performed in the ICR cell followed by three data-dependent MS/MS events in the ion trap at an isolation width of 2 m/z and normalized collision energy of 30% for 30 ms; 1+ charge states were rejected from MS/MS and dynamic exclusion was employed for 45 s after 2 repeat analyses in a 30 s window.

Bottom-up data analysis. RAW LC-MS files were processed using MASCOT Distiller (Matrix Science, Boston, MA, USA) to create peak lists in .mgf format to be searched against the IPI human database using MASCOT (Matrix Science, London, UK) [43]. Searches employed a peptide mass tolerance of ±5 ppm and a fragment ion tolerance of ±1 Da. Fixed modification included carboxyamidomethylation of cysteine, with variable modifications including methionine oxidation and deamidation of asparagines and glutamine residues.

Results and Discussion

H9 cells cultured in CM and SILAC CM for at least five passages were lysed, combined at a 1:1 ratio as determined by Bradford assay, and subjected to an initial sample cleaning procedure with C8 reversed-phase

column packing material. The resulting SILAC mixture was analyzed by nano-flow LC-MS/MS as described vide supra. Figure 1 shows examples of two proteins identified by top-down LC-MS/MS. Figure 1a demonstrates the detection of an 8.5 kDa SILAC pair with charge states ranging from 6+ to 12+. The light form of the 7+ charge state was isolated for collision induced dissociation, allowing the protein to be identified as ubiquitin c splice variant using the biomarker search method in ProSightPC. The sequence of the protein, identified with an E value of 2.94E-19, was then used to calculate a mass difference between the light and heavy protein forms of 80.1798 Da based on the presence of four arginine residues and seven lysine residues. This mass difference translates to an m/z difference of 11.4542 at $z = 7$, which strongly correlates with the mass difference between the SILAC pair in the experimental precursor spectrum. This increases the confidence of the protein identification.

Likewise, the principle of making an initial identification from MS and MS/MS data and confirming with mass difference of the SILAC pair is also illustrated in Figure 1b. Here, SILAC pairs of a 12.3 kDa protein with charge states ranging from 9+ to 13+ are detected, with the light form of the 12+ charge state being isolated for MS/MS analysis. Searching in absolute mass mode with ProSightPC identified the protein as macrophage migration inhibitory factor with an E value of 3.02E-7. Although the noisy broadband and MS/MS spectra, in addition to only a few identified product ions, would normally cast doubt on the confidence of the identification, the calculated mass difference for five arginines and three lysines of 54.1432 Da between the light and heavy forms given by the predicted sequence again matches the m/z difference of 4.5119 at $z = 12$ of the experimental precursor ion spectrum, giving additional confidence to the identified sequence. We were similarly able to identify 11 intact proteins from hESC lysate (Table 1). For each protein, the mass difference between the SILAC pair peaks was matched to the number of labels in the identified sequence within a tolerance of 10 ppm. Seven of the 11 identified proteins were full intact proteins identified by the absolute mass search in ProSightPC, and four were identified by biomarker search, meaning they are truncated forms of intact protein sequences. Four of the proteins identified by top-down analysis are variants of thymosin or prothymosin, of which only one form was identified in a parallel bottom-up study (data not shown). Five of the 11 proteins identified were acetylated at the N-terminus, including histone H4. Interestingly, three of the proteins identified in Table 1 returned multiple protein hits with E values less than the cutoff value of E-4 when searched and could not be unambiguously identified using ProSight PC 2.0. These included the ubiquitin c splice variant (five proteins), nucleophosmin 1 isoform 3 (four proteins), and hypothetical protein LOC140699 isoform 1 (six proteins), and are shown in Table 2 along with other protein forms that met the cutoff value of
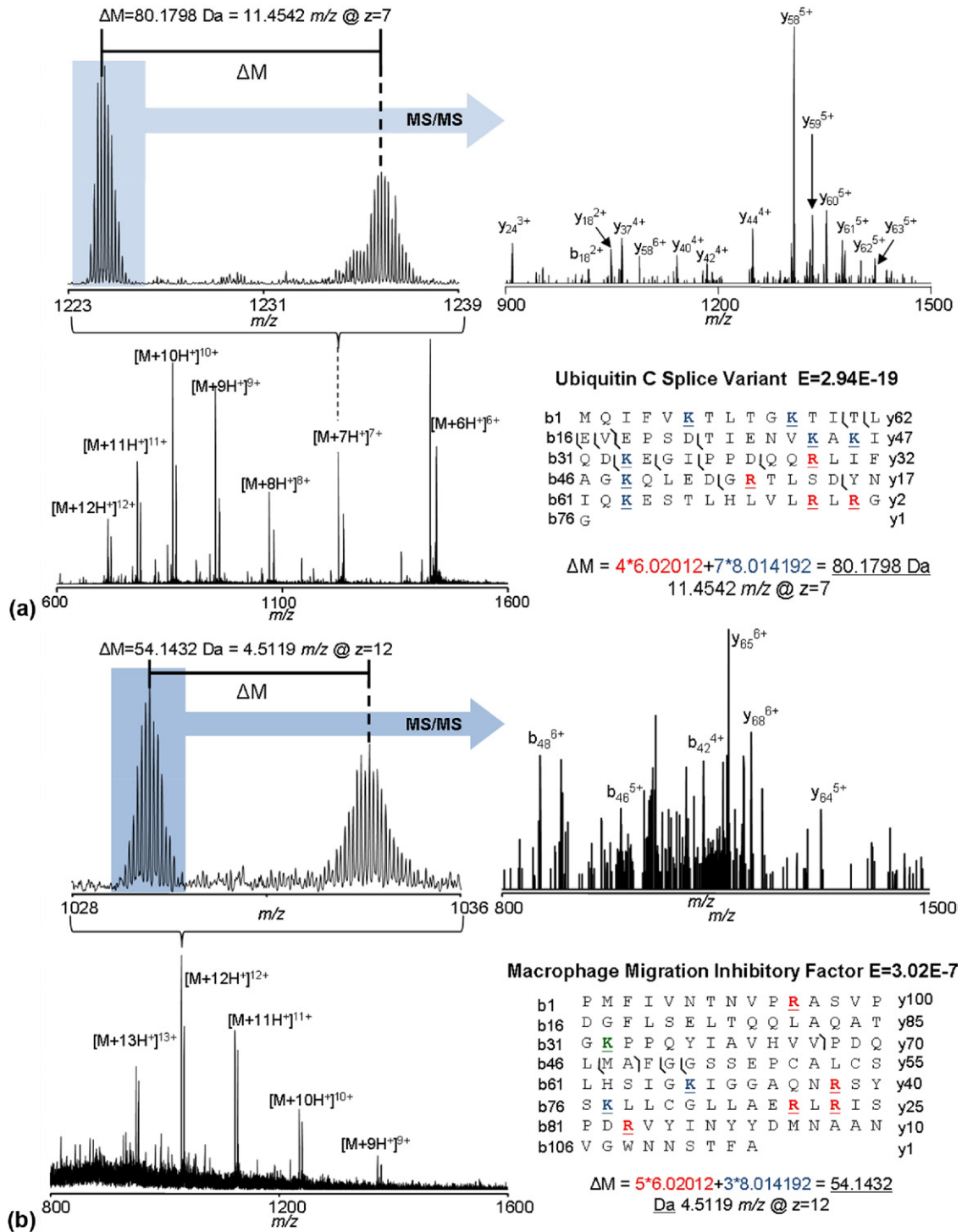
Figure 1. Precursor and product ion mass spectra of intact protein SILAC pairs are shown in addition to their identified sequences and E-values as scored by ProSightPC for (**a**) ubiquitin c splice variant and (**b**) macrophage migration inhibitory factor.

1E-4. Analyzing the mass difference between the SILAC pairs for each of the corresponding spectra and comparing with the number of arginine and lysine residues in each case yielded unambiguous protein identification. For example, nucleophosmin 1 isoform 3 containing 1 Arg and 3 Lys was identified with an E value of 5.88E-9 in a biomarker search in ProSight PC in addition

to three other proteins that met the cutoff value for identification of 1E-4. These included neurofascin (E = 5.94E-5) containing 1 Arg and 1 Lys, ubiquitin carboxyl-terminal hydrolase (E = 1.16E-6) containing 1 Lys, and DC30 (E = 3.30E-6) containing 2 Arg. Analyzing the mass difference between the light and heavy protein signal gave a ΔM of 30.0627 Da, which correlates to one

Table 1. hESC proteins identified by top-down quantitative proteomic analysis

Accession number	Protein description	Theoretical mass (Da)	Observed mass (Da)	Mass diff. (ppm)	No. matching fragments	E value	P score	Search type
IPI00455510	Isoform 2 of prothymosin α	11977.90	11977.79	−9.18	18	4.67E-24	1.17E-29	Absolute mass
IPI00220828	Thymosin β-4	4960.49	4960.46	−6.80	35	1.84E-46	4.6E-52	Absolute mass
IPI00658013	Nucleophosmin 1 isoform 3	3096.51	3096.53	0.53	6	5.88E-09	3.58E-14	Biomarker
IPI00180240	Thymosin β-4-like protein 3	4970.51	4970.49	−3.22	9	1.59E-9	3.97E-15	Absolute mass
IPI00329028	WD Repeat-containing protein 5B	3864.04	3864.04	−0.88	8	6.31E-9	3.84E-14	Biomarker
IPI00793729	Ubiquitin C splice variant	8559.62	8559.62	−0.43	14	2.94E-19	1.79E-24	Biomarker
IPI00220827	Thymosin β-10	4933.52	4933.52	−1.01	29	4.15E-46	1.04E-51	Absolute mass
IPI00220362	10 kDa heat shock protein, mitochondrial	10835.80	10835.75	−4.60	13	2.05E-12	5.13E-18	Absolute mass
IPI00218635	Hypothetical protein LOC140699 isoform 1	4528.44	4528.44	0.38	5	4.24E-05	2.58E-10	Biomarker
IPI00293276	Macrophage migration inhibitory factor	12337.20	12337.29	7.30	6	3.02E-7	6.22E-13	Absolute mass
IPI00453473	Histone H4	11388.40	11388.33	−6.15	7	1.38E-5	3.47E-10	Absolute mass

labeled arginine and three labeled lysine residues. Of the four total protein sequences that met the E value cutoff, only nucleophosmin contained labeled residues that matching the composition given by the ΔM of the experimental spectrum.

Although heavy and light pairs of SILAC-labeled intact proteins were detected, the ratio of abundance of light to heavy protein initially appeared to deviate from the expected 1:1 ratio. A closer inspection of the isotopic distributions of the heavy labeled proteins, illustrated in the upper left panels of both Figure 1a and b, reveal more complexity to the distribution than is present in the light protein signal, with additional isotopic peaks appearing at both lesser and greater *m/z* values than expected. This complexity is a compounded effect of the incorporation efficiency of the two different stable-isotope labeled amino acids, in addition to contribution from the metabolic conversion of heavy arginine to heavy proline [38, 44]. However, the SILAC ratio calculated as the sum of light isotope abundances divided by the heavy isotope abundances still results in a value of unity. Specifically, the SILAC ratios calculated for ubiquitin c splice variant and macrophage migration inhibitory factor were 1.05 and 0.90, respectively, both falling within the range of variability of SILAC ratios observed in bottom-up SILAC experiments with the same samples (data not shown).

In addition to the ability to confirm protein identifications with the ΔM of the light and heavy forms of the protein observed in the precursor ion spectra, the incorporation of amino acid labels in MS/MS fragment ions of the same protein can be confirmed in cases where both the heavy and light forms of the protein are selected for CID. Figure 2 shows an overlay of the

Table 2. Intact protein identifications facilitated by amino acid label counting

Accession number	Protein description	Theoretical mass (Da)	Observed mass (Da)	Mass diff. (ppm)	No. matching fragments	Amino acid labels	E value
IPI00793729	Ubiquitin c splice variant	8559.62	8559.62	−0.43	14	7K 4R	2.94E-19
IPI00793330	12 kDa protein	8559.62	8559.62	−0.42	8	7K 3R	1.87E-12
IPI00186004	Ankyrin repeat domain 36B	8559.55	8559.62	7.62	8	5K 5R	1.41E-12
IPI00935575	Hypothetical protein XP_002342919	8559.53	8559.62	9.80	7	2K 5R	8.23E-07
IPI00022557	Proteinase-activated receptor 4	8559.54	8559.62	8.6015	6	0K 3R	4.21E-06
IPI00658013	Nucleophosmin 1 isoform 3	3096.51	3096.53	0.53	6	1R 3K	5.88E-09
IPI00470575	Neurofascin isoform 9	3096.54	3096.53	−8.74	4	1R 1K	5.94E-05
IPI00184533	Ubiquitin carboxyl-terminal hydrolase	3096.49	3096.53	6.69	4	0R 1K	1.16e-06
IPI00022947	DC30	3096.53	3096.53	−5.99	4	2R 0K	3.30e-06
IPI00218635	Hypothetical protein LOC140699 isoform 1	4528.44	4528.44	0.38	5	4K 2R	4.24E-05
IPI00885141	Isoform 3 of FAM149A	4528.47	4528.44	−6.33	4	3K 3R	6.71E-05
IPI00011465	Thymic stromal cotransporter homolog	4528.42	4528.44	6.49	4	3K 3R	8.19E-05
IPI00885173	Isoform 4 of FAM149A	4528.47	4528.44	−6.33	4	4K 3R	1.26E-05
IPI00878107	Highly similar to interleukin 6	4528.48	4528.44	−8.48	4	4K 1R	6.82E-05
IPI00027547	Dermcidin	4828.48	4528.44	−6.56	4	7K 0R	9.21E-05

MS/MS spectra of the ubiquitin c splice variant shown also in Figure 1. Varying differences in m/z between the same y and b ion peaks arise from each fragment containing a different number of labeled amino acids. By comparing the difference in mass between those peaks with the mass difference predicted by the assigned sequence of the protein, we can further increase our confidence as to the identity of the protein analyte. For instance, the y_{18}^{2+} ion are expected to contain one lysine and two arginines that should be heavier in the labeled form of the protein by 20.054 Da. The m/z difference of 10.027 in the experimental spectra ($z = 2$)

shown in Figure 1 corresponds closely to the predicted mass difference. Likewise, the light and heavy forms of the y_{58}^{5+} ion have a predicted mass difference 64.15 Da predicted by the assigned sequence, containing five lysines and four arginines. The 12.83 m/z difference at $z = 5$ from the overlaid experimental spectra also match with the predicted mass difference, and further confirm the identification of the protein sequence.

Because the stable isotope labeled amino acids were not 100% purely labeled, there exists an increasing statistical probability that one or more light amino acids will be incorporated as the number of amino acids to be
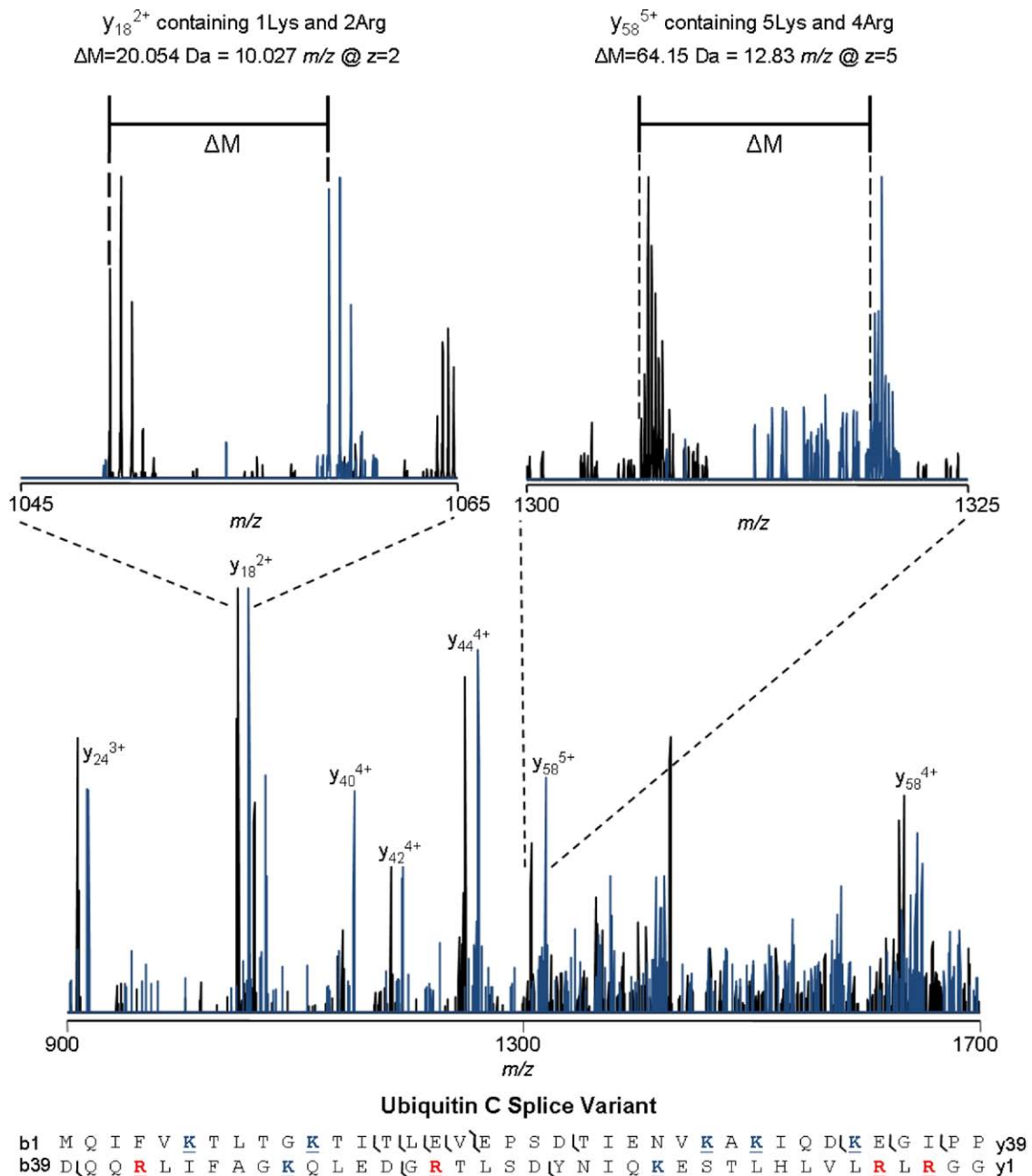

Figure 2. The MS/MS spectra of light (black) and heavy (blue) forms of ubiquitin c splice variant are overlaid to demonstrate the confirmation of fragment amino acid sequences based on their differences in mass as correlated to the different numbers of labeling amino acids incorporated.

labeled in the protein increases. The result of such a probability is an incomplete labeled intact protein, being slightly lower in mass than the completely labeled counterpart. The abundance of these unlabeled forms is modeled in eq 1, which takes into account the number of amino acids to be labeled in the protein, N, and the incorporation efficiency of the amino acid label, E, and determines the relative abundance of the protein signal arising from n unlabeled residues, A_n [29].

$$A_n = \frac{N!}{n!\,(N-n)!} E^n \,(1-E)^{N-n} \tag{1}$$

To account for the use of both stable isotope labeled arginine and lysine, in addition to the conversion of heavy arginine to heavy proline, the abundance of the completely labeled isotopic distribution is described by eq 2:

$$A_H = A_{100\%} - A_{100\%}\left[\left[1 - \left[\frac{N_R!}{n_R!\,(N_R - n_R)!}E_R^{n_R}\right.\right.\right.$$
$$\left.\left.\times (1-E_R)^{N_R - n_R}\right]\right] + \left[1 - \left[\frac{N_K!}{n_K!\,(N_K - n_K)!}E_K^{n_K}\right.\right.$$
$$\left.\left.\times (1-E_K)^{N_K - n_K}\right]\right] + N_p\frac{\%P}{100}\right] \tag{2}$$

Where A_H the abundance of the completely labeled protein, $A_{100\%}$ is the abundance if labels were incorporated 100% of the time, N_R is the number of arginine residues in the protein, n_R is the number of unlabeled arginine residues in the protein, and E_R is the incorporation efficiency of arginine. N_K, n_K, and E_K represent the same parameters for stable isotope labeled lysine, and N_P and %P are the number of prolines and percent of conversion from arginine to proline, respectively.

To test the model described in eq 2 with the isotopic distributions observed in the top-down analysis of SILAC labeled hESCs, the amino acid incorporation efficiency of both Arg6 and Lys8 were determined from a parallel bottom-up proteomic analysis of just the heavy-labeled cell lysate. Figure 3 shows results for peptides corresponding to proteins identified in Figure 1, including the ubiquitin c splice variant (Figure 1a, Figure 3a, b) and macrophage migration inhibitory factor (Figure 1b, Figure 3c, d). Figure 3a shows the mass spectrum of the peptide identified by MS/MS as ESTLHLVL*R* from the ubiquitin c splice variant, including the dominant signal from the form of the peptide containing Arg6, and the much weaker signal from the light form, amplified 50 times in the inset. Extracting the ion chromatograms of the monoisotopic masses from both the light and heavy forms and finding the ratio of their peak areas suggests arginine incorporation to be ~98%. Similarly, Figure 3b shows the calculated incorporation of Lys8, also ~98%, from the mass spec-

trum and extracted ion chromatogram of the peptide TLSDYNIQ*K*, which corresponds to the ubiquitin c splice variant as well.

Peptides that correspond to macrophage migration inhibitory factor are illustrated in Figure 3c and d. Both were arginine terminated, and yielded Arg6 incorporation values of ~98.5%. The contribution arginine-proline conversion is visible in the mass spectrum of the peptide identified as PMFIVNTNVP*R* (Figure 3c) as indicated by the increased abundance of the fifth and sixth isotopic peaks in the labeled peptide signal. The area of the extracted ion chromatogram of the fifth isotopic peak, which is actually the monoisotopic peak of the heavy proline-containing form of the peptide, compared with that of the heavy label, reveals roughly 5% conversion of arginine to proline.

Using the percent incorporation values of Arg6 and Lys8 in addition to the percent conversion of Arg6 to Pro5 obtained from bottom-up proteomic analysis, eqs 1 and 2 can be used in combination to plot the predicted abundance of a completely SILAC labeled intact protein signal (using eq 2) in addition to signals that arise from having one or more unlabeled lysine or arginine residue and furthermore the signal arising from the incorporation of Pro5 (using eq 1). Figure 4 shows the experimental mass spectra of two intact proteins overlaid upon the inverted theoretical spectra as predicted by eqs 1 and 2. The experimental spectrum of the 7+ charge state of the ubiquitin c splice variant is illustrated in Figure 4a, where the predictive model accounts for 98% incorporation of Lys8 and Arg6 as determined by the peptide level data discussed above and shown in Figure 3a, b. Five percent conversion of arginine to proline is also taken into account in the predicted spectra. All of the theoretical predicted spectra are displayed on an inverted axis with the sum of the abundances of each isotopic peak overlaid upon the experimental spectrum, showing a strong correlation between the two. The theoretical isotopic distribution also strongly correlates with the experimental spectrum for the macrophage migration inhibitory factor SILAC pair shown in Figure 3b.

Comparing the heavy isotopic distribution of Figure 4a to that of Figure 4b, the statistical relationships of the model become more apparent. For instance, the maximum abundance of the heavy form of the ubiquitin c splice variant is just over half that of the light form, whereas the heavy form of the macrophage migration inhibitory factor is ~60% that of the light form. The abundance of the SILAC labeled protein signal to its unlabeled counterpart is related to the number of labeling amino acids present for which imperfect incorporation subtracts from the abundance of the completely labeled form. This is evident in that the ubiquitin c splice variant has 11 total arginine and lysine residues for labeling versus eight residues for macrophage migration inhibitory factor. The greater number of labeling amino acids in ubiquitin increases the probability of
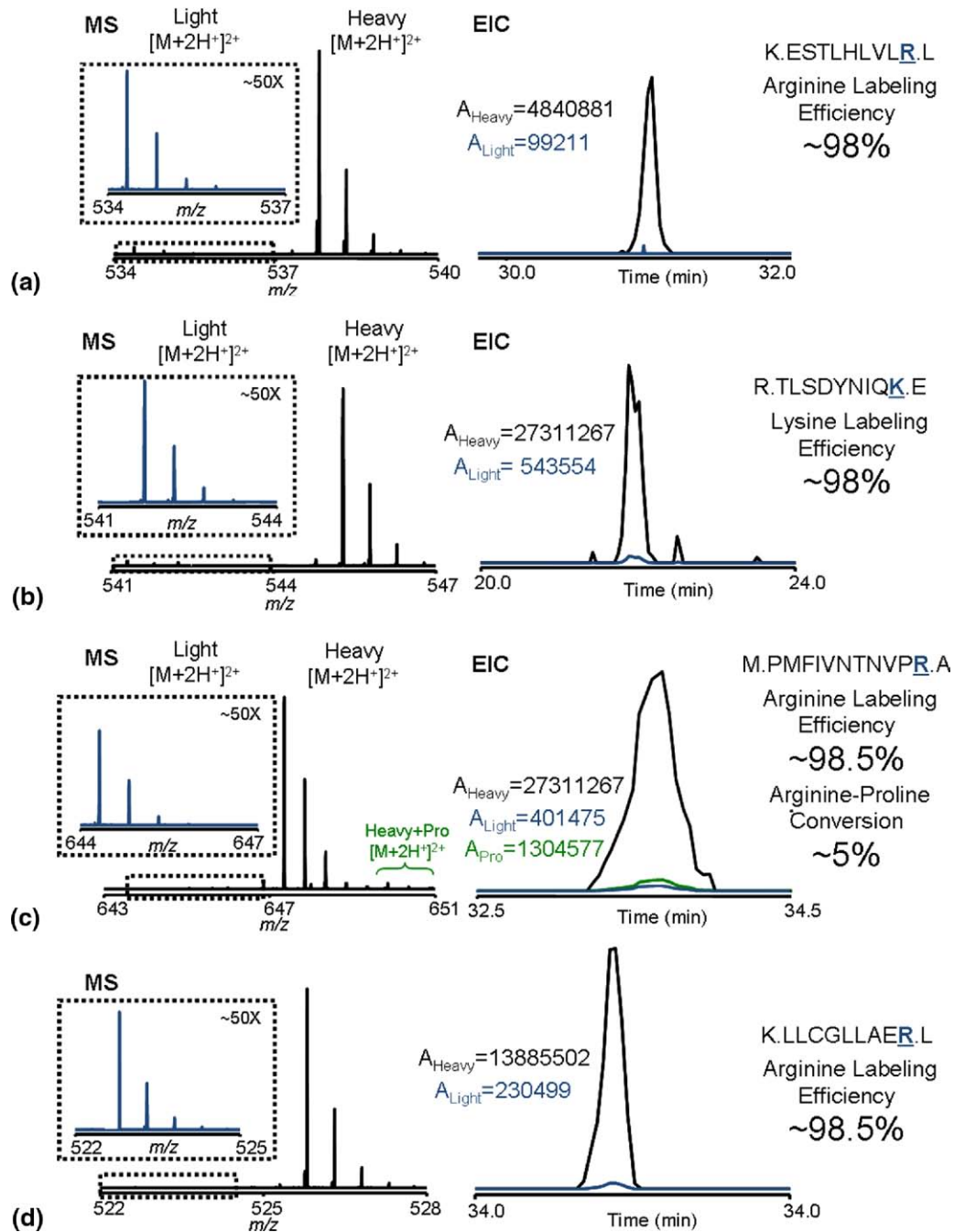
Figure 3. MS spectra and extracted ion chromatograms of light and heavy of arginine and lysine terminated peptides are used to calculate the percent incorporation of each amino acid label into (**a**) and (**b**) ubiquitin c splice variant, and (**c**) and (**d**) macrophage migration inhibitory factor. Proline conversion is also demonstrated (**c**).

one being unlabeled, leading to a more pronounced "shoulder" of incompletely labeled protein in Figure 4a compared with Figure 4b. Additionally, two proline residues are present for potential incorporation of heavy proline in ubiquitin c splice variant and eight proline residues are present in macrophage migration inhibitory factor, which in turn lead to a slightly more pronounced "tail" from the heavy proline protein signal in the latter relative to the former.

Conclusions

Quantifying intact proteins using the SILAC method remains a challenging task, even with the relatively high incorporation of stable isotope labeled lysine and arginine < 98%. While the use of dual labeling strategies is beneficial in the case of bottom-up quantification by SILAC, even the small amount of light amino acid incorporated in the place of the stable isotope label is
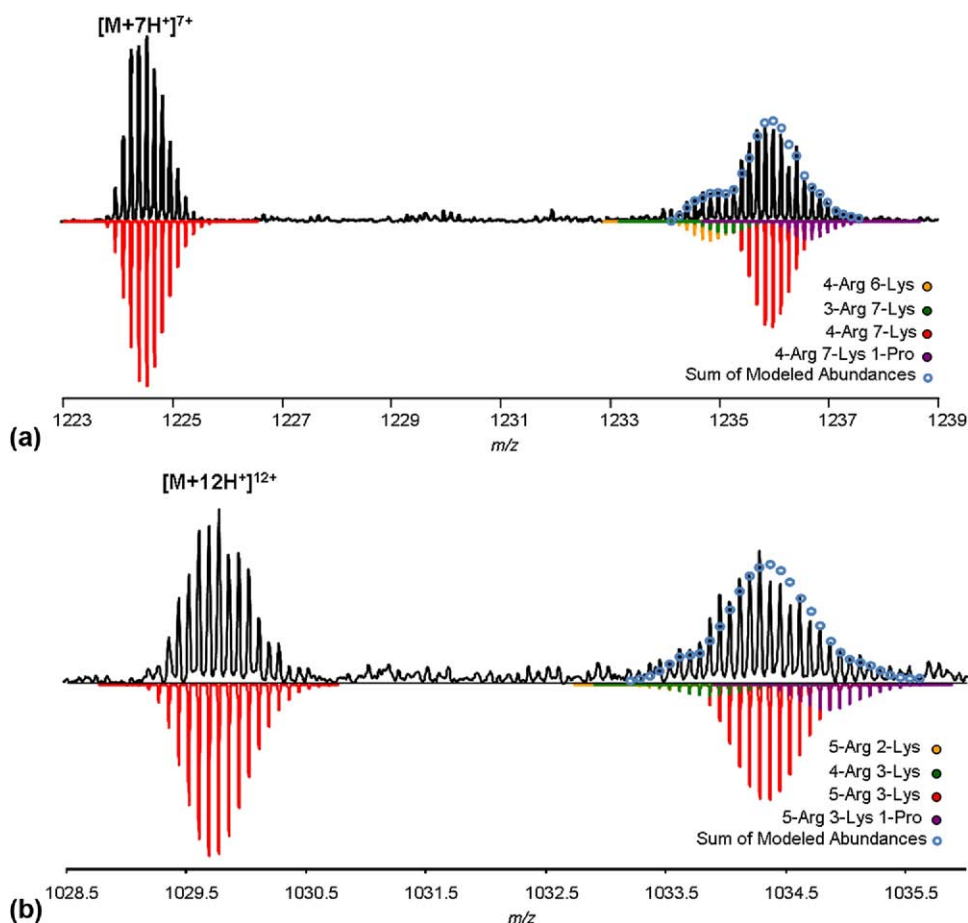
Figure 4. Experimental spectra are shown along with model spectra as predicted by eqs 1 and 2. The sum of model isotopic abundances of the inverted model spectra (circles) are superimposed on the experimental spectra of (**a**) ubiquitin c splice variant and (**b**) macrophage migration inhibitory factor.

compounded. This effect is magnified further when a labeling amino acid is converted by the biological system into a different residue, as is the case with arginine-proline conversion. Many of these effects can be taken into account easily with quantification software with bottom-up experiments, but remain unaddressed from the top-down perspective. Importantly, multiple intact proteins were unambiguously identified, including multiple intact forms of proteins for which only a single form was identified from bottom-up analysis. It would be beneficial in future studies involving intact proteins for the use of a single, highly purified, essential amino acid label to be utilized for SILAC labeling. Stable isotope labeled lysine would be a strong candidate for future use, as it occurs frequently enough in the proteome to be present in most proteins, is a limiting amino acid, and is not converted biologically into other amino acids. The result of this labeling system would be less complex heavy protein isotope distributions for easier intact protein quantification, giving researchers the ability to thoroughly examine the intact proteomes of human embryonic stem cells in addition other biological systems.

Acknowledgments

References

1. Vallier, L.; Pedersen, R. A. Human Embryonic Stem Cells: An In Vitro Model to Study Mechanisms Controlling Pluripotency in Early Mammalian Development. *Stem Cell Rev.* **2005,** *1,* 119–130.
2. Kelleher, N. L. Top-Down Proteomics. *Anal. Chem.* **2004,** *76*(11), 196A–203A.
3. Bogdanov, B.; Smith, R. D. Proteomics by FTICR Mass Spectrometry: Top Down and Bottom Up. *Mass Spectrom. Rev.* **2005,** *24*(2), 168–200.
4. Kelleher, N. L.; Taylor, S. V.; Grannis, D.; Kinsland, C.; Chiu, H. J.; Begley, T. P.; McLafferty, F. W. Efficient Sequence Analysis of the Six Gene Products (7–74 kDa) from the *Escherichia coli* Thiamin Biosynthetic Operon by Tandem High-Resolution Mass Spectrometry. *Protein Sci.* **1998,** *7*(8), 1796–1801.
5. Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W. Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry. *J. Am. Chem. Soc.* **2002,** *124*(4), 672–678.
6. Meng, F. Y.; Du, Y.; Miller, L. M.; Patrie, S. M.; Robinson, D. E.; Kelleher, N. L. Molecular-Level Description of Proteins from Saccharomyces cerevisiae Using Quadrupole FT Hybrid Mass Spectrometry for Top Down Proteomics. *Anal. Chem.* **2004,** *76*(10), 2852–2858.
7. Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-Down Mass Spectrometry. *Nat. Methods* **2007,** *4*(10), 817–821.
8. Dillon, T. M.; Bondarenko, P. V.; Ricci, M. S. Development of an Analytical Reversed-Phase High-Performance Liquid Chromatograph Electrospray Ionization Mass Spectrometry Method for Characteriza-

tion of Recombinant Antibodies. *J. Chromatogr. A* **2004,** *1053*(1053), 299–305.

9. Tran, J. C.; Doucette, A. A. Multiplexed Size Separation of Intact Proteins in Solution Phase for Mass Spectrometry. *Anal. Chem.* **2009,** *81*(15), 6201–6209.

10. Tran, J. C.; Doucette, A. A. Gel-Eluted Liquid Fraction Entrapment Electrophoresis: An Electrophoretic Method for Broad Molecular Weight Range Proteome Separation. *Anal. Chem.* **2008,** *80*(5), 1568–1573.

11. Wang, Y.; Balgley, B. M.; Rudnick, P. A.; Lee, C. S. Effects of Chromatography Conditions on Intact Protein Separations for Top-Down Proteomics. *J. Chromatogr. A* **2005,** *1073*, 35–41.

12. Syka, J. E. P.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F. Novel Lnear Quadrupole Ion Trap/FT Mass Spectrometer: Performance Characterization and Use in the Comparative Analysis of Histone H3 Post-Translational Modifications. *J. Proteome Res.* **2004,** *3*(3), 621–626.

13. Hu, Q. Z.; Noll, R. J.; Li, H. Y.; Makarov, A.; Hardman, M.; Cooks, R. G. The Orbitrap: A New Mass Spectrometer. *J. Mass Spectrom.* **2005,** *40*(4), 430–433.

14. Yates, J. R.; Cociorva, D.; Liao, L. J.; Zabrouskov, V. Performance of a Linear Ion Trap-Orbitrap Hybrid for Peptide Analysis. *Anal. Chem.* **2006,** *78*(2), 493–500.

15. Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: Improved Protein Identification and Characterization for Top Down Mass Spectrometry. *Nucleic Acids Res.* **2007,** *35*, W701–W706.

16. Pesavento, J. J.; Bullock, C. R.; LeDuc, R. D.; Mizzen, C. A.; Kelleher, N. L. Combinatorial Modification of Human Histone H4 Quantitated by Two-dimensional Liquid Chromatography Coupled with Top-Down Mass Spectrometry. *J. Biol. Chem.* **2008,** *283*(22), 1427–1437.

17. Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L. A Robust Two-Dimensional Separation for Top-Down Tandem Mass Spectrometry of the Low-Mass Proteome. *J. Am. Soc. Mass Spectrom* **2009,** *20*(12), 2183–2191.

18. Solis, R. S.; Ge, Y.; Walker, J. W.; Single Amino. Acid Sequence Polymorphisms in Rat Cardiac Troponin Revealed by Top-Down Tandem Mass Spectrometry. *J. Muscle Res. Cell Motil.* **2008,** *29*, 203–212.

19. Wu, S.; Lourette, N. M.; Tolic, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Pasa-Tolic, L. An Integrated Top-Down and Bottom-Up Strategy for Broadly Characterizing Protein Isoforms and Modifications. *J. Proteome Res.* **2009,** *8*, 1347–1357.

20. Gordon, E. F.; Mansoori, B. A.; Carroll, C. F.; Muddiman, D. C. Hydropathic Influences on the Quantification of Equine Heart Cytochrome *c* Using Relative Ion Abundance Measurements by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *J. Mass Spectrom.* **1999,** *34*(10), 1055–1062.

21. Martinovic, S.; Veenstra, T. D.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. Selective Incorporation Acids for Identification Proteome-Wide Level. *J. Mass Spectrom.* **2002,** *37*(1), 99–107.

22. Veenstra, T. D.; Martinovic, S.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. Proteome Analysis Using Selective Incorporation of Isotopically Labeled Amino Acids. *J. Am. Soc. Mass Spectrom.* **2000,** *11*(1), 78–82.

23. Julka, S.; Regnier, F. Quantification in Proteomics Through Stable Isotope Coding: A review. *J. Proteome Res.* **2004,** *3*(3), 350–363.

24. Zhang, R. J.; Sioma, C. S.; Wang, S. H.; Regnier, F. E. Fractionation of Isotopically Labeled Peptides in Quantitative Proteomics. *Anal. Chem.* **2001,** *73*(21), 5142–5149.

25. Du, Y.; Parks, B. A.; Sohn, S.; Kwast, K. E.; Kelleher, N. L. Top-Down Approaches for Measuring Expression Ratios of Intact Yeast Proteins Using Fourier Transform Mass Spectrometry. *Anal. Chem.* **2006,** *78*(3), 686–694.

26. Ong, S.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteom.* **2002,** *1*(5), 376–386.

27. Ong, S.; Kratchmarova, I.; Mann, M. Properties of 13C-Substituted Arginine in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). *J. Proteome Res.* **2003,** *2*, 173–181.

28. Waanders, L. F.; Hanke, S.; Mann, M. Top-Down Quantitation and Characterization of SILAC-Labeled Proteins. *J. Am. Soc. Mass Spectrom.* **2007,** *18*(11), 2058–2064.

29. Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. A.; Muddiman, D. C. Top-Down Identification and Quantification of Stable Isotope Labeled Proteins from Aspergillus flavus Using On-Line Nano-Flow Reversed-Phase Liquid Chromatography Coupled to a LTQ-FTICR Mass Spectrometer. *Anal. Chem.* **2008,** *80*(13), 4994–5001.

30. Choi, M. Y.; An, Y. J.; Kim, S. H.; Roh, S. H.; Ju, H. K.; Hon, S. S.; Park, J. H.; Cho, K. J.; Cho, D. W.; Kwon, S. W. Mass Spectrometry Based Proteomic Analysis of Human Stem Cells: A Brief Review. *Exp. Mol. Med.* **2007,** *39*(6), 690–695.

31. Swaney, D.; Wenger, C. D.; Thomson, J. A.; Coon, J. J. Human Embryonic Stem Cell Phosphoproteome Revealed by Electron Transfer Dissociation Tandem Mass Spectrometry. *Proc. Nat. Acad. Sci. U.S.A.* **2009,** *106*(4), 995–1000.

32. Van Hoof, D.; Munoz, J.; Braam, S. R.; Pinske, M. W.; Linding, R.; Heck, A. J.; Mummery, C. L.; Krijgsveld, J. Phosphorylation Dynamics During Early Differentiation of Human Embryonic Stem Cells. *Cell Stem Cell* **2009,** *5*, 214–226.

33. Phanstiel, D.; Brumbaugh, J.; Berggren, W. T.; Conard, K.; Feng, X.; Levenstein, M. E.; McAlister, G. C.; Thomson, J. A.; Coon, J. J. Mass Spectrometry Identifies and Quantifies 74 Unique Histone H4 Isoforms in Differentiating Human Embryonic Stem Cells. *Proc. Nat. Acad. Sci. U.S.A.* **2008,** *105*(11), 4093–4098.

34. Williamson, A. J. K.; Smith, D. L.; Blinco, D.; Unwin, R. D.; Pearson, S.; Wilson, C.; Miller, C.; Lancashire, L.; Lacaud, G.; Kouskoff, V.; Wheeton, A. D. Quantitative Proteomics Analysis Demonstrates Post-transcriptional Regulation of Embryonic Stem Cell Differentiation to Hematopoiesis. *Mol. Cell. Proteom.* **2008,** *7*(3), 459–472.

35. Graumann, J.; Hubner, N. C.; Kim, J. B.; Ko, K.; Moser, M.; Kumar, C.; Cox, J.; Scholer, H.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Proteome Quantitation of Mouse Embryonic Stem Cells to a Depth of 5111 Proteins. *Mol. Cell. Proteom.* **2008,** *7*(4), 672–683.

36. Prokhorova, T. A.; Rigbolt, K. T.; Johansen, P. T.; Henningsen, J.; Kratchmarova, I.; Kassem, M.; Blagoev, B. Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Quantitative Comparison of the Membrane Proteomes of Self-Renewing and Differentiating Human Embryonic Stem Cells. *Mol. Cell. Proteom.* **2009,** *8*, 959–970.

37. Xu, C.; Inokuma, M. S.; Denham, J.; Golds, K.; Kundu, P.; Gold, J. D.; Carpenter, M. K. Feeder-Free Growth of Undifferentiated Human Embryonic Stem Cells. *Nat. Biotechnol.* **2001,** *19*(10), 971–4.

38. Van Hoof, D.; Pinkse, M. W.; Oostwaard, D. W.; Mummery, C. L.; Heck, A. J.; Krijgsveld, J. An Experimental Correction for Arginine to Proline Conversion Artifacts in SILAC-Based Quantitative Proteomics. *Nat. Methods* **2007,** *4*, 677–678.

39. Price, P. J.; Goldsborough, M. D., Tilkin, M. L. Embryonic Stem Cell Serum Replacement. International patent application WO98/30679, 1998.

40. Winston, R. L.; Fitzgerald, M. C. Concentration and Desalting of Protein Samples for Mass Spectrometry Analysis. *Anal. Biochem.* **1998,** *262*, 83–85.

41. Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Nat. Acad. Sci. U.S.A.* **1996,** *93*(25), 14440–14445.

42. Andrews, G. L.; Shuford, C. M.; Burnett, J. C.; Hawkridge, A. M.; Muddiman, D. C. Coupling of a Vented Column with Splitless NanoRPLC-ESI-MS for the Improved Separation and Detection of Brain Natriuretic Peptide-32 and Its Proteolytic Peptides. *J. Chrom. B* **2009,** *877*(10), 948–954.

43. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry data. *Electrophoresis* **1999,** *20*(18), 3551–3567.

44. Bendall, S. C.; Hughes, C.; Stewart, M. H.; Doble, B.; Bhatia, M.; Lajoie, G. A. Prevention of Amino Acid Conversion in SILAC Experiments with Embryonic Stem Cells. *Mol. Cell. Proteom.* **2008,** *7*(9), 1587–1597.