DOI 10.1002/pmic.201300438 *Proteomics* 2014, *14*, 1130–1140

REPORT

The first pilot project of the consortium for top-down proteomics: A status report

Xibei Dang[1,2], Jenna Scotcher[1], Si Wu[3], Rosalie K. Chu[3], Nikola Tolić[3], Ioanna Ntai[4], Paul M. Thomas[4], Ryan T. Fellers[4], Bryan P. Early[4], Yupeng Zheng[4], Kenneth R. Durbin[4], Richard D. LeDuc[5], Jeremy J. Wolff[6], Christopher J. Thompson[6], Jingxi Pan[7], Jun Han[7], Jared B. Shaw[8], Joseph P. Salisbury[9], Michael Easterling[6], Christoph H. Borchers[7], Jennifer S. Brodbelt[8], Jeffery N. Agar[9], Ljiljana Paša-Tolić[3], Neil L. Kelleher[4] and Nicolas L. Young[1]*

[1] Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL, USA
[2] Department of Chemistry and Biochemistry, Florida State University, Tallahassee, FL, USA
[3] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA
[4] Departments of Chemistry and Molecular Biosciences and the Proteomics Center of Excellence, Northwestern University, Evanston, IL, USA
[5] NIH/NCRR Mass Spectrometry Resource, Washington University in St. Louis, St. Louis, MO, USA
[6] Bruker Daltonics, Billerica, MA, USA
[7] UVic-Genome BC Proteomics Centre, Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada
[8] Department of Chemistry, University of Texas, Austin, TX, USA
[9] Departments of Chemistry and Pharm. Sci., Barnett Institute, Northeastern University, Boston, MA, USA

Pilot Project #1—the identification and characterization of human histone H4 proteoforms by top-down MS—is the first project launched by the Consortium for Top-Down Proteomics (CTDP) to refine and validate top-down MS. Within the initial results from seven participating laboratories, all reported the probability-based identification of human histone H4 (UniProt accession P62805) with expectation values ranging from 10^{-13} to 10^{-105}. Regarding characterization, a total of 74 proteoforms were reported, with 21 done so unambiguously; one new PTM, K79ac, was identified. Inter-laboratory comparison reveals aspects of the results that are consistent, such as the localization of individual PTMs and binary combinations, while other aspects are more variable, such as the accurate characterization of low-abundance proteoforms harboring >2 PTMs. An open-access tool and discussion of proteoform scoring are included, along with a description of general challenges that lie ahead including improved proteoform separations prior to mass spectrometric analysis, better instrumentation performance, and software development.

Received: October 1, 2013
Revised: February 25, 2014
Accepted: March 13, 2014

Keywords:
Human histone H4 / PTM analysis / Technology / Top-down proteomics

Additional supporting information may be found in the online version of this article at the publisher's web-site

Correspondence: Dr. Nicolas Young, Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Florida State University, 1800 E. Paul Dirac Drive, Tallahassee, FL 32310, USA
E-mail: nyoung@magnet.fsu.edu
Fax: 850-644-1366

Abbreviations: **CTDP**, Consortium for Top-Down Proteomics; **ECD**, electron capture dissociation; **ETD**, electron transfer dissociation; **FT-ICR**, Fourier transform-ion cyclotron resonance; **FTMS**, Fourier transform MS; **PUF**, ProSight Upload Format; **QCAD**, quadrupole selected collisional activation; **UVPD**, ultraviolet photodissociation

*Current Address: Cardiovascular Division, King's College London, The Rayne Institute, St. Thomas' Hospital, London, SE1 7EH, UK

1 Introduction

Here, we report preliminary results of the first pilot project of the Consortium for Top-down Proteomics. The Consortium for Top Down Proteomics (CTDP) was founded in 2012 with the aim of promoting innovative research, collaboration, education, and accelerating the comprehensive analysis of intact proteins (http://www.topdownproteomics.org). To validate top-down MS as a viable and effective tool for defining the proteome and increasing awareness of top-down techniques in the wider proteomics community, CTDP launched Pilot Project #1. This first pilot project focuses on the identification and characterization of human histone H4 by top-down MS [1–3].

The aim of Pilot Project #1 is to establish the level to which multiple laboratories can identify and localize PTMs and ultimately define the same proteoforms in a complex mixture. This is regardless of the top-down techniques employed. A proteoform is a recently introduced term to fully describe all sources of heterogeneity possible originating from a single gene [4]. Importantly for this work, this includes the combinations of PTMs that may arise from multiple sites of variable PTM. Each specific combination of PTMs is a separate proteoform. For example, five variable sites of phosphorylation may generate as many as 2^5 or 32 proteoforms. Human histone H4, which contains 102 amino acids, was chosen as the target because of its high abundance; easy access; high profile involvement, and significance in epigenetic regulation and maintenance; and more importantly, its heavily modified nature. Histone H4 usually contains 1–6 modifications on a single molecule. Therefore, it is challenging to identify and localize its PTMs and even more challenging to simultaneously determine all PTM localizations to fully characterize a proteoform [5–8].

Human histone H4 samples from HeLa S3 cells were purified, pooled into a single homogeneous sample, aliquoted, and then distributed equally to multiple laboratories. Each laboratory performed top-down MS and data analysis independently to identify histone H4 and characterize all proteoforms present. Each laboratory, with its unique combination of ion source, fragmentation method, mass analyzer, and data analysis system, reported their identified proteoforms, with or without ambiguity in PTM localization to the CTDP. PTM localization and proteoform characterization information was reported by a total of seven laboratories and analyzed for interlaboratory comparison.

2 Materials and methods

2.1 Histone H4 preparation

HeLa S3 suspension cells were maintained in Joklik's modified MEM media and harvested after reaching 10^7 cells/mL density. Cell nuclei were isolated by standard nuclei isolation with nuclei isolation buffer (15 mM Tris HCl, 60 mM KCl, 15 mM NaCl, 5 mm $MgCl_2$, 1 mM $CaCl_2$, 250 mM sucrose with 0.3% NP-40. There are also 1 mM dithiothreitol, 10 nM microcystin, 0.5 mM 4-(2-aminoethyl) benzenesulfonyl fluoride, and 10 mM sodium butyrate added as inhibitors.) Histones were extracted from the isolated nuclei through standard acid extraction by 0.2 M H_2SO_4 on a rotator at 4°C for 2 h. Insoluble material was removed after centrifugation at 3400 × *g* and histone was precipitated from the supernatant with 20% trichloroacetic acid (final, w/v) on ice for 45 min. Histone extracts was air-dried after acetone wash [9].

Histone extracts were further purified and separated by RP-HPLC using Thermo Ultimate 3000 system (Thermo Fisher Scientific, San Jose, CA, USA) on a Vydac C18 column (218TP54, 250 × 4.6mm, Grace, Deerfield, IL, USA). The gradient was a linear 30% B to 60% B gradient with 0.3% B/min increase at 0.8 mL/min flow rate (Buffer A = 5% ACN:water with 0.2% TFA; Buffer B = 95% ACN:water with 0.188% TFA). Histone H4 eluted at ~44% B, collected by an autocollector and dried under vacuum. The H4 fractions were then suspended in water, combined and aliquoted into equal portions. Each aliquot contained ~25 μg of histone H4 extracted from ~5×10^7 cells.

2.2 Online LC separation

Two laboratories performed online LC as described below.

(1) Agilent 1200 nanoflow pumps (Agilent Technologies), 2-position Valco valves (Valco Instruments Co., Houston, TX, USA), and a PAL autosampler (Leap Technologies, Carrboro, NC, USA) were assembled into custom HPLC system allowing for fully automated histone analysis using in-house packed WAX-HILIC column (PolyCAT A, 75 μm id X 100 cm, 5 um particles, 1000 Å pore size, PolyLC, Columbia, MD, USA). Mobile phases consisted of 70% ACN aqueous solution with 1.0% formic acid (A) and 70% ACN with 12% formic acid (B). The aliquoted human histone H4 sample was resuspended in water to a final concentration of 0.2 μg/μL. Te microliters of the diluted sample was first loaded onto an SPE column (150 μm id × 5 cm, 5 um particles, 1000 Å pore size) for 10 min using buffer A, and the following LC gradient with a flow rate of 300 nL/min was applied for proteoform separation: 0–1 min, 0–10% B; 1–10 min, 10–35% B; 10–200 min, 35–90% B; 200–207 min, 90–100% B; 207–240 min, 100% B; 242 min, 100% A. The eluted proteins were detected on-line using high-resolution FT-MS (Fourier transform MS) as described later.

(2) Reversed-phase LC was performed using an Eksigent two-dimensional LCHPLC, a self-packed 12 cm, 150 μm id column with 5 μm C18 beads (unpacked from a larger Targa column). Buffer A consisted of 0.1% formic acid v/v in HPLC grade water and buffer B consisted of 0.1%

formic acid v/v in 100% HPLC grade ACN v/v/. Samples were manually injected and eluted at 2.5 μL/min using a 32 min gradient: 5% B for 6 min, 5–40% B for 8 min, 40–100% B for 2 min, 100% B for 5 min, 100–2% B for 1 min, and 2% B for 10 min.

2.3 MS

No guidance or specifications were given to the individual laboratories as to how to perform MS other than that top-down proteoform characterization is the goal. Below are the top-down MS methods performed by seven laboratories in random order:

(1) Data acquisition was performed using the LTQ Orbitrap Velos (Thermo Fisher Scientific) with nominal resolving power of 60,000 ($m/z = 400$). Precursor ion mass spectra were collected for 500 to 2000 m/z (with automatic gain control (AGC) set to 1E6), followed by data-dependent ETD (electron transfer dissociation) MS/MS (isolation window 3 Th, reagent ion AGC 2E5, 15 ms reaction time, MSn AGC 5E5) of the top five most abundant ions. The number of micro scans for both MS and MSn was two. Dynamic exclusion was implemented with the exclusion duration of 200 s and an exclusion list size of 500. MS/MS was only performed on species with charge states greater than four.

(2) Dried histone fraction was acquired from CDTP and used without cleanup or separation. The histone fraction was dissolved in H_2O to create a stock with a concentration of approximately 8 pmol/μL. Prior to infusion by ESI, the histone sample was diluted to approximately 200 fmol/μL in 50:50:0.1 H2O:ACN:formic acid (Sigma, St. Louis, MO, USA) and infused with ESI at 2 μL/min. All MS experiments were performed on a Bruker 12 T solariX XR FTMS. The histone sample was first analyzed in MS mode to determine potential targets for MS/MS. For ECD, ETD, and quadrupole selected collisional activation (QCAD) experiments, precursor ions were isolated in the external quadrupole and stored in the adjacent collision cell. For QCAD, ions were dissociated inside the external collision cell adjacent to the quadrupole. For ECD, q-isolated precursor ions were transferred to the ICR cell and irradiated with 0.8 eV electrons for <50 ms. For ETD, q-isolated precursor ions were reacted with the ETD reagent for 60 ms in the collision cell. For ETD and ECD, 300 scans were averaged for each mass spectrum. ESI experiments were externally calibrated with NaTFA clusters.

(3) The histone sample was resuspended in MeOH: water:formic acid (50:49:1) at concentration of 0.1 ug/uL and analyzed by direct infusion without any prior separation or manipulation. Data were obtained on an Orbitrap Elite (Thermo Fisher Scientific) by static electrospray. The optimal spray was obtained with 2 kV applied to a New Objective glass tip with 3AP coating. For the precursor measurement, a SIM scan (range of 653.5–678.5 m/z) was used with 10 microscans and an AGC target of 1×10^5. Fragmentation was performed in a data-dependent fashion (top 3) with ETD and an AGC target of 1×10^6. The isolation width was 4 m/z and the ETD activation time was 15 ms. Dynamic exclusion was enabled with a repeat count of 2, exclusion duration of 600 s. Each MS^2 spectrum was obtained with 50 microscans. Precursor scans were analyzed with a resolution of 240 000 (at m/z 400) and MS/MS scans were performed with 120 000 resolution.

(4) All top-down ECD data were recorded on a 12 T Apex-Qe hybrid Fourier transform-ion cyclotron resonance (FT-ICR) mass spectrometer (Bruker Daltonics, Billerica, MA, USA) equipped with a nano-ESI source, a quadruple mass filter, and a hexapole collision cell. The ESI source was operated in positive ion mode at a capillary voltage of 3500 V with a spray shield voltage of 3200 V. ECD was employed as the gas-phase fragmentation method. The ECD parameters are set as follows: electron pulse length 18 ms, electron beam bias 1.1 V, grid potential 12 V, heater current 1.2 A. All ECD data were acquired from front end quadrupole-isolated [M + 14H]14+ ions. The FWHM of the transmitted isotope distribution window corresponded to 1–2 m/z units. Mass calibration was performed with ECD fragments of bovine ubiquitin.

(5) Samples were introduced via a nanospray ion source (CaptiveSpray) with a dual ion funnel (solariX) 12.0 tesla hybrid quadrupole FT-ICR, FT-MS mass spectrometer (Bruker Daltonics). Important instrument operation parameters include dry gas flow rate 4.0 L/min., neb gas flow rate 1.0 bar, capillary voltage 1555.0 V, source declustering potential = 34 V, source accumulation time = 0.001 s, ion accumulation time = 0.1 s, TOF = 0.001 ms, and sidekick extraction voltages = −1.5 V.

(6) Twenty-five micrograms histone H4 sample was resuspended in 100 μL water and further dilute ten times in 50% water: ACN with 1% formic acid. All spectra were collected in custom-built 9.4T FT-ICR mass spectrometer. Histone was introduced into mass spectrometer through a positive nano-electrospray source at 0.4 μL/min under 2100 V voltage. Intact proteins were first filtered through a quadrupole with 5 m/z window and furthered isolated to 1 m/z window by stored waveform inverse Fourier transform. Electron capture dissociation was performed for 50 mS with 12 V electron grid potential. Each spectrum was an average of 500–1500 scans.

(7) Protein solutions were prepared at 10 μM in 49.5:49.5:1 v/v/v water: ACN: formic acid, and were analyzed by direct infusion at a flow rate of 3 μL/min into a Thermo Scientific Orbitrap Elite mass spectrometer (Bremen, Germany). The Orbitrap mass spectrometer was

modified for photodissociation by addition of a Coherent Excistar ArF excimer laser, and all ultraviolet photodissociation (UVPD) experiments were undertaken in the HCD cell (3). Spectra were acquired using a mass range of 200–2000 m/z and resolving power of 240 000 at m/z 400. The AGC target for MS^2 was set to one million, and an isolation width of 25 m/z was used. For UVPD experiments, precursor ions were transferred to HCD cell with a normalized collision energy of 1% (no collisional induced dissociation occurs at this setting), and spectra were acquired using one laser pulse at 2 mJ/pulse at 193 nm. The HCD collision gas pressure was reduced to a pressure measured as a delta of 0.1E-10 Torr in the UHV portion of the vacuum chamber containing the Orbitrap analyzer (5 mTorr collision gas pressure). Two hundred scans were averaged per spectrum.

2.4 Data analysis

No guidance or specifications were given to the individual laboratories as to how to analyze MS/MS data. Below are the data analysis methods of the seven laboratories in random order.

(1) Tandem mass spectra were searched against a custom-built H4_HUMAN database using ProSightPC V2.0 software (Thermo Fisher Scientific). This database consisted of about 2.5 million of H4 proteoforms calculated based on the information from Uniprot. All spectra were first deconvoluted using Xtract mode embedded in ProSightPC then searched in absolute mode (precursor mass tolerance was set at 3 Da, fragment mass tolerance was set at 20 ppm, and delta mass option was enabled). Initial H4 proteoform identifications were filtered by a minimal of ten matched fragment ions. Filtered spectra were then evaluated using in-house developed functions in a semimanual process [10].
To obtain list of other proteins present in sample, we used MSAlign+ (http://bix.ucsd.edu/projects/msalign/) software on human database derived from UniProt FASTA download.

(2) For each dissociation method, monoisotopic m/z and charge state were determined using the SNAP II peak picking algorithm in Bruker Daltonics DataAnalysis. Singly charged peaks were exported to Bruker Daltonics BioTools. Using BioTools, the singly charged product ions were matched against known histone H4 sequence that contained no modifications. First, using the mass difference between the product ions, BioTools calculates long strings of amino acids called sequence tags. Second, these sequence tags are compared against the known Histone H4 sequence to determine possible matches. Third, the mass difference of the matched sequence tag versus the known Histone H4 sequence was calculated and used to determine possible modifications. Finally, the Histone H4 sequence was modified with the possible modification(s) and matched to the experimental data. "Correct" Histone modifications were manually determined by a combination of mass accuracy and sequence coverage. Matched product ion mass accuracies were <2 ppm for ETD and QCAD experiments and <3 ppm for ECD experiments.

(3) .RAW files were processed with the cRAWler algorithm inside ProSightPC 3.0. Averaged scans were assigned precursor and fragment masses by cRAWler via the Xtract algorithm. cRAWLer grouped one precursor mass to the observed fragment masses from each summed unit to create individual ProSightPC experiments, which were concatenated into an XML file in ProSight Upload Format (PUF). PUF files were then searched by ProSightPC 3.0 against a database created for human H4 with all possible combinatorial modifications. The results were manually examined to determine the ambiguous assignments from unambiguous ones for the localization of modifications based on the coverage of fragment ions and nature of modifications (e.g. acetylation only occurs at lysine).

(4) Bruker Compass Dataanalysis (version 4.0) was used to generate peak lists and identify fragment ions. ProteinProspector was used for database searching using the sequence of human histone H4. Intact ion tolerance is 10 ppm, and fragment ion tolerance is 5 ppm. Manual validation was conducted for all of the identified isoforms.

(5) Intact protein masses were reconstructed using Maximum Entropy Deconvolution from DataAnalysis (Bruker Daltonics, version 3.4) and spectra were assigned manually.

(6) Data analysis was performed using a custom ProSight PC workflow. First, the ProSight absolute mass search algorithm was modified to match the nine canonical ion types for UVPD [3, 11]. A shotgun annotated database of Histone H4 and H2A was then created to search against. Neutral masses were inferred from the averaged MS^2 spectrum acquired above using the Xtract algorithm from Thermo Fisher Scientific with a S/N cut-off of 3. These neutral masses were used to create a PUF input file that was sent to the modified algorithm for processing. Search results were manually validated.

(7) Data analysis was performed using custom in-house software (currently in development) in conjunction with manual interpretation and validation. Briefly, a basis set of every potential H4 proteoform and the resulting isotopic distributions of all ions were enumerated in silico. For each MS^2 spectrum all basis sets (proteoforms) were tested by querying with 3 ppm accuracy for all of the theoretically most intense peaks of every isotopic cluster, filtering for those with appropriate isotopic distributions. The sum of the intensities was maximized to identify the most likely candidate, explaining the largest portion of the data. Further manual validation was performed to confirm correct peak assignments and reasonable overall interpretation of the data.

3 Results and discussion

3.1 Analysis of the preliminary results: Protein identification

All seven laboratories participating reported probability-based identification of human histone H4 (UniProt accession **P62805**) with a median E-value of 10^{-37}. Individual scores ranged from 10^{-13} to 10^{-105} and some of the laboratories also identified minor contaminants of histone H2A.Z, H2A2C, H2A3, H2A2A etc. with a E-value range from 10^{-12} to 10^{-45}. Laboratories were not instructed whether to report on proteins other than histone H4 in this pilot project.

3.2 PTMs and proteoforms

There were 13 PTMs observed across all laboratories at 11 sites. On a purely mathematical basis these PTMs generate as many as 4096 proteoforms; however, it is reasonably established that not all of these are likely present due to biological specificity and certainly not in equal abundance. In this work, there were a total of 74 proteoforms identified by the seven laboratories, 21 of which are without ambiguity. This number of proteoforms is remarkably similar to previous middle down analyses of H4 [6, 7]. Among all unambiguously assigned proteoforms, three of them—H4Nα-acK16acK20ac, H4Nα-acK20me2K31ac, H4Nα-acK20me2K79ac—are novel (with assigned spectra in Supporting Information). One new PTM (K79ac) is identified. This PTM is located after amino acid 23 and therefore would not have been identified by the middle-down approaches previously used to extensively characterize histone H4 proteoforms [6, 7]. The middle down efforts used AspN protease and purified the 1–23 AA peptide before analysis and thus such deep PTMs were not observed either isolated or within a proteoform.

The CTDP encouraged investigators to report proteoforms with ambiguity rather than force the localization with insufficient evidence. With high resolving power MS, ambiguity arises from PTM localization rather than degree of modification. In Pilot Project #1, 53 ambiguous proteoforms were reported. Ambiguity mostly results from insufficient sequence coverage and low S/N in the MS^2 spectra but most ambiguous proteoforms still contain unambiguous PTM localization information for some PTMs. For example, one H4 proteoform was identified with four acetylations and two methylations by intact mass. The MS^2 spectrum localized two acetylations on the N-terminus and at K5 (lysine 5), two methylations on K20, and two acetylations between 31 and 59 AA. There are three locations possible for the two unassigned acetylations, K31, K44, and K59. With insufficient sequence coverage, proper localization of these two acetylations was not possible and resulted in one ambiguous proteoform. However, there are three unambiguously localized PTMs in this ambiguous proteoform: Nα-ac, K5ac, and K20me2. Only such unambiguously localized PTMs from ambiguous proteoforms, as well as all PTMs from unambiguous proteoforms (the individual PTMs of which are inherently unambiguous), were considered in our interlaboratory comparison (Fig. 1, Fig. 2, and Table 1).

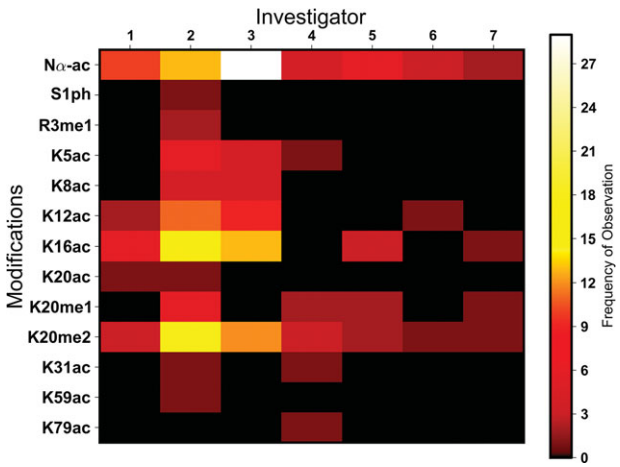

Figure 1. Heat map of the individual PTMs marginalized from the collective top-down dataset organized by investigator. The scale of the color scheme is linear to the number of observations of proteoforms that contain each PTM, offset by 1. (0 = no observations, 1 = one observation, maximum = 29 observations) Only unambiguously localized PTMs are considered.

Much of the variance observed in the interlaboratory results derives from differences in the depth and thoroughness of data analysis. Some laboratories identified more than 30 proteoforms while others identified fewer than 10 (out of 74 in total reported by the consortium as a whole). Thus, a high degree of similarity on a percentage basis is not possible. Although quantitation was not included in the goals of the pilot project there is a clear relationship between abundance and likelihood of observation. As expected, most laboratories reported the more abundant proteoforms and these dominate the overlap between laboratories. The less abundant proteoforms were only observed by laboratories performing a deeper analysis, most often employing online LC-MS approaches. Six proteoforms—H4Nα-ac, H4Nα-acK20me1, H4Nα-acK20me2, H4Nα-acK16ac, H4Nα-acK16acK20me1, H4Nα-acK16acK20me2—were reported by more than one laboratory. These correlate very well with the most abundant proteoforms according to previous semiquantitative middle-down analyses [6, 7]. Among them, the H4Nα-acK20me2 proteoform, is reported by six of seven laboratories due to its high abundance. Other proteoforms—such as H4K20me2 and H4K5acK12acK16acK20me2—are only observed by one laboratory.

When these data are marginalized to individual PTMs, as would typically be reported by bottom-up methods, a fairly uniform and consistent picture emerges as can be seen in Fig. 1. Marginalization is the process by which some axes of data may be projected onto the remaining axes, via integration. In this case, it is achieved simply by summing the number of observations of proteoforms that contain each

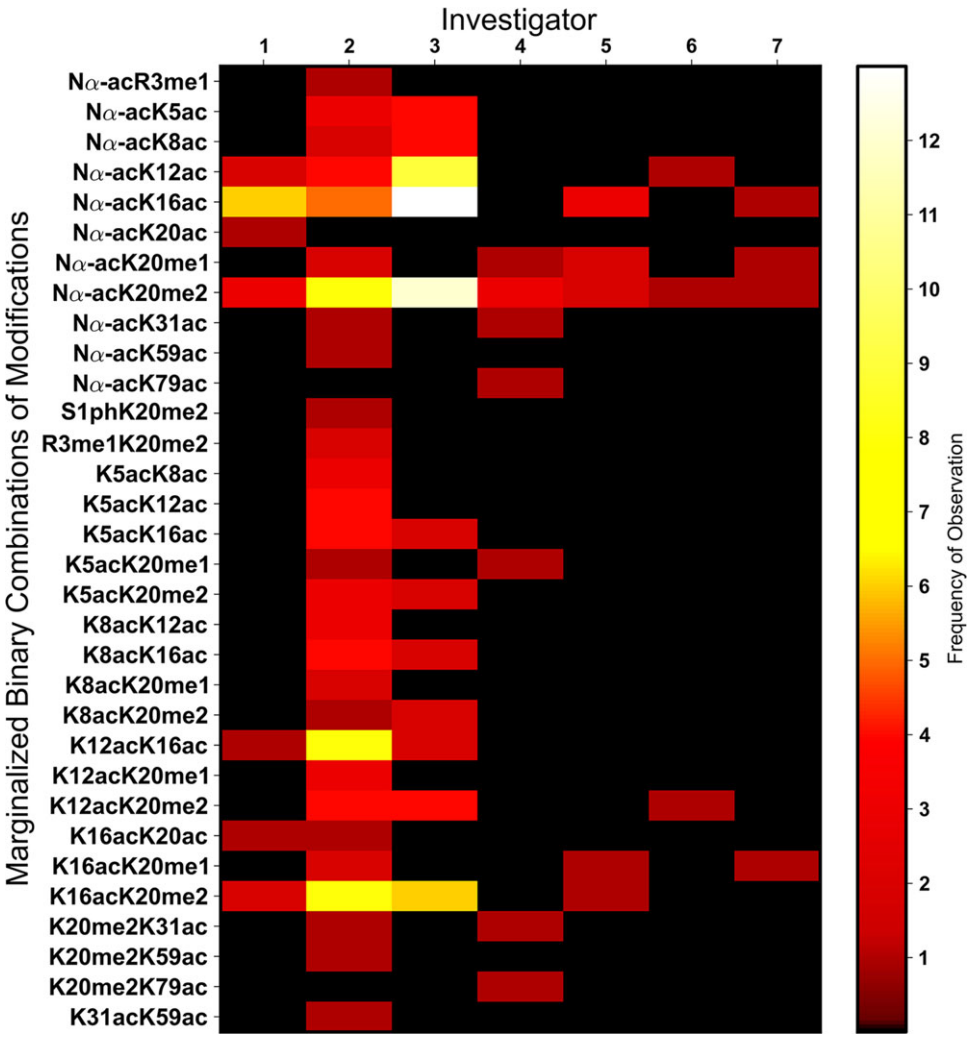

Figure 2. Heat map of the binary combinations of PTMs marginalized from the collective top-down dataset organized by investigator. This marginalization is achieved by summing the number of observations of proteoforms that contain both PTMs. Note that the presence of a binary combination of PTMs in the figure does not mean that there is a proteoform that contains only these two PTMs to the exclusion of other PTMs. The scale of the color scheme is linear to the number of observations of proteoforms that contain each binary combination of PTMs, offset by 1. (0 = no observations, 1 = one observation, maximum = 13 observations) Only unambiguously localized PTMs are considered.

PTM. Note that the presence of a PTM in the figure does not mean that there is a proteoform that contains this and only this PTM. Nα-ac, K16ac, and K20me2 are the most commonly observed PTMs in the various proteoforms reported by all laboratories. Consistency lies in the fact that the more frequent a given PTM is observed in the laboratories reporting a large number of proteoforms, such as laboratories 2 and 3, the more likely it will be reported by other laboratories. This reflects primarily the relative abundance of these PTMs. Laboratories may report similar PTMs on average but slightly different fully characterized proteoforms due either to observation of similar but distinct proteoforms or errors in PTM localization at one or more out of multiple sites. Yet, they may still be in agreement as to the PTMs' existence and frequency of observation. Some PTMs, such as K31ac or K79ac, are only observed by one or two laboratories. In such cases, it is difficult to determine based on frequency of observation alone if these are less abundant PTMs or are erroneous localizations.

Taking this analysis a step further the results can be marginalized to binary combinations as in Fig. 2. Note that these binary combinations are not proteoforms, but represent co-occurrence of two PTMs within proteoforms. Combinations that stand out include: Nα-acK16ac, Nα-acK20me2, K16acK20me2, which is not surprising given that these are combinations of the three most frequently observed PTMs. Also the K12ac variations on these themes appear as a second tier of frequently observed binary combinations, such as Nα-acK12ac, K12acK20me2, and to a lesser extent K12acK16ac. The binary combinations of PTMs more frequently observed in the laboratories performing deeper analyses (and more proteoforms overall) are again more likely to be observed by the laboratories reporting fewer proteoforms overall.

3.3 The inherent challenge of assigning confidence in proteoform identity

As we compiled data and compared the interlaboratory results it became increasingly obvious that a proper comparison requires a more complete and rigorous metric of confidence in

Table 1. Complete table of proteoforms reported, including ambiguous proteoforms and PTM localizations

#	m.e.	a.me	a.ac	a.ph	Unambiguous PTMs	N[α]	S1	R3	K5	K8	K12	K16	K20	K31	K59	K79
1	2	0	0	0	K20me2											
4	3	0	0	0	Nα-ac											
1	4	0	0	0	K5acK20me1											
3	4	0	0	0	Nα-acK20me1											
1	4	1	0	0	Nα-ac											
1	4	1	0	0	Nα-ac											
1	4	0	1	0	K20me1											
1	4	0	1	0	K20me1											
6	5	0	0	0	Nα-acK20me2											
1	5	0	0	0	K5acK20me2											
1	6	0	0	0	Nα-acR3meK20me2											
1	6	0	0	0	Nα-acK12ac											
2	6	0	0	0	Nα-acK16ac											
1	6	0	1	0	Nα-ac											
1	6	0	1	0	Nα-ac											
1	6	0	1	0	R3meR20me2											
1	6	0	2	0	(none)											
2	7	0	0	0	Nα-acK16acK20me1											
1	7	1	0	0	Nα-acK12ac											
1	7	1	0	0	Nα-acK16ac											
1	7	1	0	0	Nα-acK16ac											
1	7	0	1	0	K12acK20me1											
1	8	0	0	0	Nα-acK12acK20me2											
4	8	0	0	0	Nα-acK16acK20me2											
1	8	0	0	0	Nα-acK20me2K31ac											
1	8	0	0	0	Nα-acK20me2K79ac											
1	8	0	1	0	Nα-acK20me2											
1	8	0	1	0	K12acK20me2											
1	8	0	1	0	K16acK20me2											
1	8	0	1	0	K16acK20me2											
1	9	0	0	0	Nα-acK5acK12ac											
1	9	0	0	0	Nα-acK12acK16ac											
1	9	0	0	0	Nα-acK16acK20ac											
1	9	1	0	0	Nα-acK16acK20me2											
1	9	0	1	0	Nα-acK12ac											
1	9	0	1	0	Nα-acK16ac											
1	9	0	1	0	Nα-acK16ac											
1	9	0	2	0	Nα-ac											
1	9	0	2	0	K16ac											
1	9	0	3	0	(none)											
1	10	1	0	0	Nα-acK5acK16ac											
1	10	1	0	0	Nα-acK8acK16ac											
1	10	1	0	0	Nα-acK12acK16ac											
1	10.7	0	1	0	S1phK20me2											
1	10.7	2	0	1	Nα-ac											
1	11	0	0	0	K5acK8acK16acK20me2											
1	11	0	0	0	K5acK12acK16acK20me2											
1	11	0	1	0	Nα-acK8acK20me2											
1	11	0	1	0	Nα-acK12acK20me2											
1	11	0	1	0	Nα-acK12acK20me2											
1	11	0	1	0	Nα-acK16acK20me2											
1	11	0	1	0	Nα-acK16acK20me2											
1	11	2	1	0	K12acK16ac											
1	11	0	2	0	Nα-acK20me2											
1	11.7	0	0	1	Nα-acK5ac											
1	11.7	0	0	1	Nα-acK8ac											
1	11.7	0	0	1	Nα-acK12ac											
1	12	1	0	0	Nα-acK20me2K31acK59ac											
1	12	0	2	0	Nα-acK16ac											
1	12	0	2	0	Nα-acK12ac											
1	12	0	2	0	K12acK16ac											
1	13	0	1	0	K5acK8acK16acK20me1											
1	14	0	1	0	Nα-acK5acK16acK20me2											
1	14	0	1	0	Nα-acK8acK16acK20me2											
1	14	0	1	0	Nα-acK12acK16acK20me2											
1	14	0	1	0	Nα-acK12acK16acK20me2											
1	14	0	2	0	Nα-acK5acK20me2											
1	14	0	2	0	Nα-acK12acK20me2											
1	14	0	2	0	Nα-acK16acK20me2											
1	14	0	2	0	K12acK16acK20me2											
1	15	0	0	0	Nα-acK5acK8acK12acK16ac											
1	15	0	3	0	K16acK20ac											
1	16	0	0	0	Nα-acK5acK8acK12acK16acK20me1											
1	17	2	5	0	(none)											

Light gray: acetylation, dark gray: 1–3 methylation, stripes: phosphorylation. Abbreviations: # = number of labortories that identified the same proteoform; m.e. = methyl equivalence, the total number of methylations required to achieve the modified mass, regardless of the actual PTMs present (e.g. 1 ac = 3 me); a.me = ambiguously assigned methylation; a.ac = ambiguously assigned acetylation; a.ph = ambiguously assigned phosphorylation; Nα-ac = N-terminal acetylation.

identity of the proteoforms. Such statistical models are nonexistent to underdeveloped and definitely not well established or widely accepted. Recent efforts in the analysis of top-down and related middle-down data while successful in identifying proteoforms within a spectrum or LC-MS run [12–14] have only made at most initial attempts at addressing the issue of robust confidence scoring across different analyses or laboratories.

The statistics of confidence in top-down data are affected by issues beyond S/N and sequence coverage. A typical Top-Down MS experiment is demonstrated in Fig. 3A. The degree of modification can be readily identified by intact mass, often expressed in the form of methyl equivalents (m.e.). Proteoforms with the same m.e. are often co-isolated during MS/MS. The fragment ions assigned in MS^2 provides information on PTM localization and proteoform characterization. In some cases where a single species is present within an MS1 peak isolated for fragmentation, the sources of error are more analogous to the peptide identification problem that most MS^2 algorithms and confidence scores were originally designed for. That is increased S/N results in increased sequence coverage and decreased unassigned peaks. The major difference is simply the substantially increased complexity, peak density, and an increased probability of interferences. This simple case (Case i, Fig. 3B), where proteoform characterization is the singular goal is widely familiar to the broad proteomics community. The reality of many Top-Down analyses is that there are frequently many complicating and confounding factors at play that need to be treated properly in order to understand the confidence that might be placed on a given proteoform. This has become clear to the top-down community; however, these same issues are more frequent in bottom-up analyses than has been traditionally recognized. Thus, there is wider potential impact in the proper scoring of proteoforms that takes into account the other sources of error more apparent in top-down data.

One of the cases we frequently observe in the consortium dataset is that some, but not all of the PTMs present are localized. This is demonstrated in Case ii, Fig. 3B. There is insufficient sequence coverage between G7 and K12. The second acetylation therefore cannot be localized. It may be on K12, K16, or a mixture of K12 and K16. This results in what we refer to as an ambiguous proteoform. The confidence in the protein identification is not necessarily strongly affected by this partial characterization and is at least semi-independent of PTM localization per se. Some PTMs, the K5 acetylation in this case, are clearly confidently assigned in these ambiguous proteoforms. Yet, we do not have an established statistical metric or nomenclature to define a well localized PTM on an incompletely characterized peptide/protein.

The most demonstrative case of the challenge of defining a proper statistical framework for scoring proteoform characterization is the case where multiple proteoforms are present. In Case iii in Fig. 3B, we present a mixed spectrum of 2 proteoforms, H4K5acK8ac and H4K8acK16ac. Such mixed spectra are quite common in top down proteomics and more

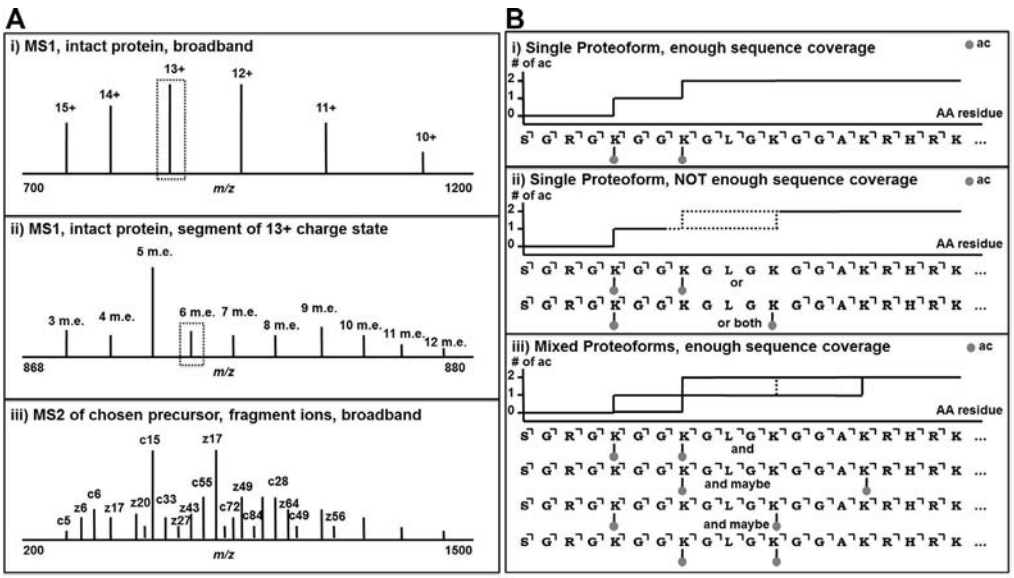

Figure 3. (A) Scheme of top-down MS experiment. (i) The broadband mass spectrum of the intact protein. (ii) Segment of the 13+ charge state of the intact protein with methyl equivalents (m.e.) labeled. The proteoform with six m.e. is selected as the precursor for the MS^2 experiment. (iii) Tandem mass spectrum for precursor chosen in (ii) to localize PTMs and identify proteoforms. (B) Ambiguity of Top-Down MS. (i) With only a single proteoform existing and enough (100%) sequence coverage, there is no ambiguity for the proteoform H4K5acK8ac. (ii) With a single proteoform existing but not enough sequence coverage, there is ambiguity on the localization of the second acetylation (K8 or K12), resulting in two ambiguous proteoforms, H4K5acK8ac or H4K5acK12ac. (iii) With mixed spectra for multiple proteoforms and enough (100%) sequence coverage, there remains uncertainty on the existence of an acetylation on K12, therefore two ambiguous proteoforms (H4K5acK12ac and H4K8acK12ac) arise.

common than is often recognized in bottom-up analyses, especially when PTM localization is involved. Two proteoforms/peptides that have the same amino acid sequence and same degree of modification are structural isomers. They tend to coelute, have the same exact mass and are thus fragmented together. In bottom-up MS this is common for phosphopeptides; however, the most abundant species is typically exclusively identified. This results in some penalty to the confidence due to other unexplained peaks. In our analysis here it is clear to us that sometimes the presence of another proteoform does not substantially affect our judgment of confidence, yet in other cases it can completely eliminate all confidence in the existence of that proteoform. In the example of case iii in Fig. 3B, even with 100% sequence coverage the presence of H4K8acK16ac and H4K5acK8ac will generate a false-positive assignment of H4K8acK12ac. A fully annotated spectrum, as required by many proteomics journals, can be generated from correctly assigned ions and supports this false assignment to current standards. The confidence in individual PTM localizations is lowered and acetylation on K12 is incorrectly assigned. This also serves as an example as to why the PTM localization score is not necessarily independent of proteoform scoring. Our experience here shows that ambiguity arises from a combination of insufficient sequence coverage (case ii, Fig. 3B) and mixed spectra (case iii, Fig 3B). Therefore, a multitiered scoring approach that provides semi-independent metrics of Protein ID, PTM localization, and proteoform characterization is needed. The Consortium for Top-Down Proteomics and its various meetings has served as a platform for discussion of these issues and we have begun work on establishing the ontology and statistical framework for such a tool.

3.4 Good scoring for good automation: A three-tiered challenge

As depicted in Fig. 4, several types of scores need to be developed and validated by the Top-Down community. PTM mapping by bottom-up is confronting similar challenges presently (e.g. in localizing phosphorylation sites for generating reliable and community-wide metrics for FDR rates). Here, Top-Down Proteomics requires an additional type of score as depicted in Fig. 4. The first tier of our scheme is a protein identification score. This metric is fairly well established; however, it should be recognized that the other tiers of proteoform characterization in our scheme may inform or detrimentally affect the confidence in the protein ID in ways not currently applied in most scoring algorithms. PTM localization is the second tier and scoring the statistical significance thereof represents a challenge even in simple cases and warrants further development. A rule of thumb for PTM localization in some laboratories has been that at least two fragment ions (on each side of the PTM and containing the PTM) are needed to report a specific PTM; however this is not formalized into a statistical model, nor standardized across laboratories. The third tier is the complete proteoform, the level at which this

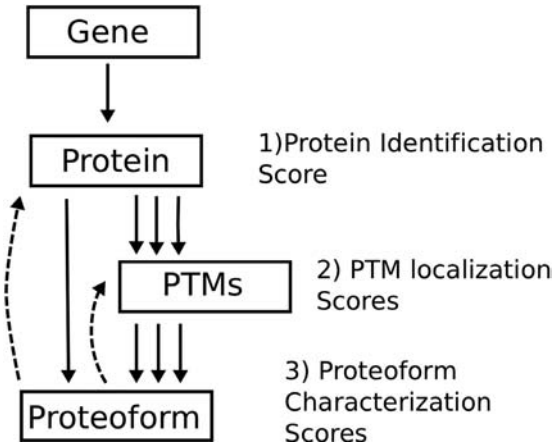

Figure 4. Depiction of three general types of scores required to capture and convey the full complexity of data generated by MS/MS of whole, multiply-modified proteins. The dashed arrows represent that these tiers of scoring are only semi-independent and each level of characterization does indeed affect and inform the other. In some cases, failure to consider all levels of information properly can result in incorrect confidence scores and misidentification.

pilot project has focused. Such proteoform level scoring of statistical significance is essentially nonexistent at this point. Clearly, agreed-to scores that convey the quality of MS/MS data and confidence in the characterization of PTMs and proteoforms in Top-Down MS experiments are needed.

The issue of proteoform and PTM scoring arose during workshops and consortium meetings and was identified as a key ambiguity in the field that needs to be addressed. For the Score Type 3 depicted in Fig. 4, an initial tool was developed for scoring proteoforms of multiply-modified proteins. The input is: (i) an intact mass, (ii) a fragment ion dataset, (iii) a protein sequence, and (iv) the specific kinds of PTMs to be considered. The output is given as a Frap score, where the higher scores are better. Initial use of the tool (available on the website of the Consortium for Top Down Proteomics) is in progress (Fig. 5). This software is ongoing in its development and represents our initial attempt to address the pressing issue of formalizing confidence in proteoform characterization. Clearly, more work on automated scoring is required by the field and we look forward to establishing community norms.

4 Concluding remarks

All participating laboratories successfully identified histone H4 as the main component of the sample provided. In total, 74 H4 proteoforms were identified by the seven laboratories, 21 of which are without ambiguity. Although proteoform-by-proteoform results reported by different laboratories indeed vary, there is great consistency buried within this variance. More abundant proteoforms were identified by the majority of the laboratories while less abundant ones were only

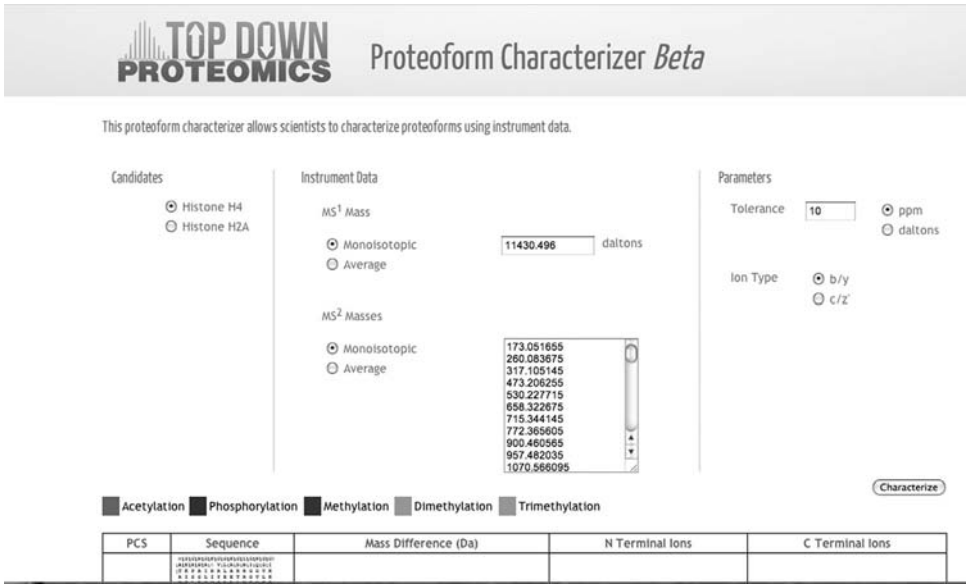

Figure 5. Screen grab from a web-based tool for generating a Proteoform Characterization Score (PCS).

identified by one or two laboratories. From the heat map of single PTMs (Fig. 1) and binary PTM combinations (Fig. 2) it is clear that the more frequently a certain PTM or binary PTM combination is reported by one laboratory, for example, K20me2 or Nα-acK20me2, the more likely it is to be reported by multiple laboratories. Therefore, the variance between laboratories has more to do with the sensitivity of methods and depth of data analysis rather than true disagreement between various top-down MS approaches.

One of the major challenges of the H4 proteoform analysis is that the various possible proteoforms are frequently structural isomers of each other. This makes it essentially impossible to separate them in the mass spectrometer. The laboratories using chromatographic methods achieved substantially more extensive lists of proteoforms, likely due to overcoming both ionization suppression and the difficulties of analyzing mixed tandem mass spectra. However, MS/MS becomes more challenging when performed at the LC timescale due to the lack of manual optimization for each MS scan and achieving a limited number of scans for each proteoform as they elute. Similarly, the relatively low resolution of the parent ion isolation often used in automated LC-MS analyses to achieve good precursor ion signal results in multiplexed MS/MS spectra due to incomplete isolation in the mass dimension.

Another limitation identified in these experiments and resulting discussions is a need for continued improvement in high-resolution isolation and high-efficiency fragmentation capabilities. Among the three common fragmentation techniques (CID/HCD, ECD, ETD), CID/HCD is less than ideal for top-down MS because of the known risk of PTM loss (mostly for phosphorylations). ECD suffers from a lower efficiency but benefits greatly from the ability to perform stored waveform inverse Fourier transform isolation in an ICR cell to completely isolate a single methyl equivalent precursor. ETD, on the other hand, is typically more efficient than ECD but high resolution isolations are not available on most ETD enabled instruments (linear ion trap hybrids). UVPD is not as common as the other three and it is still under development. However, UVPD typically produces a larger number of fragment ions of different types (a ions, x ions, b ions, y ions, c ions, and z ions) and should thus increase the confidence in proteoform assignment. Overall, increased attention to robust high-resolution isolation capabilities of future mass spectrometers in addition to efficient (on a LC time scale) and information rich fragmentation capabilities would benefit the top-down community greatly.

Another key issue for the future of top-down MS is continued improvements in data analysis with a particular need to develop universally accepted metrics of confidence. Data analysis is the most time consuming part of top down proteomics. Good software should first consider all possible proteoforms that are supported by a spectrum regardless of incomplete sequence coverage or the complexity of mixed proteoforms. Unambiguous assignment should be made only if evidence exists to confidently distinguish a proteoform (or a defined mixture of proteoforms) from all other possible explanations of the data, including potential false-positive scenarios. Current software often only picks up the most abundant proteoform, greatly limiting the depth of analysis. It is essential that a proteoform specific confidence metric be developed to evaluate the quality of each individual assignment, as discussed above in "Good Scoring for Good Automation: A Three-tiered Challenge". Unfortunately most current software can only score the proper assignment of the protein sequence but not proteoform.

Overall, it appears that analysis depth is the major source of variance between laboratories. Direct analysis of the most abundant species tends to result in high inter-laboratory concordance. Routine and robust methods capable of accessing

minor components appear to be the critical path forward for top-down proteomics.

Funding is gratefully acknowledged from NIH (R01GM067193, 1R01NS065263), NSF (CHE-1012622, JSB; ABI-1062432, RDL; DMR-11-57490), and the Welch Foundation (F-1155, JSB). The authors at the University of Victoria – Genome BC Proteomics Centre would like to thank Genome Canada and Genome BC for Science and Technology Innovation Centre funding and the Western Economic Diversification of Canada for platform support. A portion of this research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. The authors would also like to thank Northwestern University and the State of Florida.

The authors have declared no conflict of interest.

5 References

[1] Kelleher, N. L., Top-down proteomics. *Anal. Chem.* 2004, *76*, 196A–203A.

[2] Cui, W, Rohrs, H. W., Gross, M. L., Top-down mass spectrometry: Recent developments, applications and perspectives. *Analyst* 2011, *136*, 3854–3864.

[3] Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E. et al., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 2011, *480*, 254–258.

[4] Smith, L. M., Kelleher, N. L., CTDP, Proteoform: a single term describing protein complexity. *Nat. Methods* 2013, *10*, 186–187.

[5] 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*, 14927–14937.

[6] Pesavento, J. J., Mizzen, C. A., Kelleher, N. L., Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: Human histone H4. *Anal. Chem.* 2006, *78*, 4271–4280.

[7] Phanstiel, D., Brumbaugh, J., Berggren, W. T., Conard, K. et al., Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *PNAS* 2008, *105*, 4093–4098.

[8] Young, N. L., DiMaggio, P. A., Plazas-Mayorca, M. D., Baliban, R. C. et al., High Throughput Characterization of Combinatorial Histone Codes. *Mol. Cell. Proteomics* 2009, *8*, 2266–2284.

[9] Siuti, N., Roth, M. J., Mizzen, C. A., Kelleher, N. L., Pesavento, J. J. Gene-specific characterization of human histone H2B by electron capture dissociation. *J. Proteome Res.* 2006, *5*, 233–239

[10] Shen, Y., Tolic, N., Hixson, K. K., Purvine, S. O. et al., De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Anal. Chem.* 2008, *80*, 7742–54.

[11] Shaw, J. B., Li, W., Holden, D. D., Zhang, Y. et al., Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation, *J. Am. Chem. Soc.*, 2013, *135*, 12646–12651.

[12] DiMaggio, P. A. Jr., Young, N. L., Baliban, R. C., Garcia, B. A., Floudas, C. A., A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Mol. Cell. Proteomics* 2008, *8*, 2527–2543.

[13] Zamdborg, L., LeDuc, R. D., Glowacz, K. J., Kim, Y. B. et al., ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* 2007, *35*(Web Server issue), W701–W706.

[14] Liu, X., Sirotkin, Y., Shen, Y., Anderson, G. et al., Protein identification using top-down. *Mol. Cell Proteomics* 2012, *11*, M111.008524.

