Supplementary Information

Mapping Intact Protein Isoforms in Discovery Mode Using Top Down Proteomics

John C. Tran, Leonid Zamdborg, Dorothy R. Ahlf, Ji Eun Lee, Adam D. Catherman, Kenneth R. Durbin, Jeremiah D. Tipton, Adaikkalam Vellaichamy, John F. Kellie, Mingxi Li, Cong Wu, Steve M. M. Sweet, Bryan P. Early, Nertila Siuti, Richard D. LeDuc, Philip D. Compton, Paul M. Thomas, and Neil L. Kelleher*

*Corresponding author E-mail: n-kelleher@northwestern.edu

General Description and Background

Over its century of development, mass spectrometry (MS) has undergone several stages in its evolution. Initial atomic analyses[1] eventually gave way to detection of intact molecules[2] and their structural interrogation by ion fragmentation. In the past few decades, even large protein complexes have been launched into the gas phase[3], with size limits for linear protein molecules well above 2,000 amino acids[4]. Most recently, the direct detection of protein fragment ions produced in the gas phase[4] has created excitement about another major leap forward in the evolution of tandem MS[5]. Unfortunately, large scale fragmentation of intact proteins by MS has proven exceedingly difficult above ~3 kDa in size. This is the practical reason that proteases like trypsin are used to make complex mixtures even more so, despite digesting the natural isoforms[6,7,a] that correlate to and even control complex phenotypes at the molecular level[8]. Direct interrogation of intact proteins has shown the potential to overcome the "isoform problem" (see main text) for individual proteins[9-13], but has not been achieved previously on a proteome scale. This is largely attributed to the lack of robust fractionation methods for intact proteins that are well integrated with liquid chromatography and electrospray MS. Below, we describe a new separation platform which has the recovery and resolving power to afford an unprecedented level

[a] Note that IUPAC has recently recommended that isoform be used only for related protein forms that arise from gene family members with high sequence identity or other sources of genetic variation such as polymorphism. Events such as alternative splicing or post-translational modification are suggested by IUPAC to be called "protein species" (*6-7*).

of proteome coverage using intact proteins as the primary unit of measurement for high throughput proteomics. Note that the first stage of separation used here can be exchanged with a great variety of fractionation or isolation approaches, as the last three dimensions are general and well integrated.

The 4D Separation Platform

The unprecedented coverage in this study was made possible partly through the use of sharply improved separations for intact proteins (**Fig. 1**). The first dimension of the 4D separation platform entails a custom designed eight chamber solution isoelectric focusing (sIEF) device, reported in 2008[14]. Effective partitioning of 0.5-2 mg of total protein was accomplished in a timeframe analogous to chromatographic approaches (1.5 h). Typical IEF resolution is shown in **Fig. 1a** and since the process occurs in solution, intact protein precipitation is a less serious issue than that faced by popular gel-based IEF systems. Nevertheless, the spectre of protein precipitation at its isoelectric point (pI) was offset by an SDS wash protocol on each chamber after focusing, followed by pooling these washes to the corresponding fractions. The pH gradient, generated with carrier ampholytes, span the pH range 3.5-9.5, with proteins having extreme pI values focused in the anode and cathode chambers and were recovered to ensure minimum bias against proteins with a very low or high pI. Ten sIEF fractions (including the anolyte and catholyte) are available, but were normally combined into four or five fractions before simultaneous separation using the multiplexed gel-eluted liquid fraction entrapment electrophoresis (mGELFrEE) device[15,16].

The second dimension of the platform involves using a custom GELFrEE device operating in a multiplexed format[15,16], which enables parallel fractionation of all fractions collected from sIEF[14]. Here, nine fractions were collected from each gel column (~40 total fractions). An image from analytical SDS-PAGE analysis of GELFrEE fractions originating from sIEF-Fr.6 is shown at the bottom of **Fig. 1a** and **b**. Examination of these gel images reveal the resolution and peak capacity attained in this two dimensional liquid electrophoretic (2D-LE) platform. Using 12%T polyacrylamide gel tubes enhanced separation resolution between 15-40 kDa. The 2D-LE platform required 3 h of run time to complete (90 min. from sIEF, 90 min. for mGELFrEE), and the resulting fractions covered essentially the entire pI range as well as masses

between 5 kDa-110 kDa. It is important to note that the 2D-LE platform separates analogously to the classical 2D gel platform. However, in sharp contrast to 2D gels, SDS was used to wash the sIEF device and throughout the mGELFrEE experiments so high recoveries of intact proteins were achieved[14,16]. In addition, all fractions from the 2D-LE platform are recovered in solution making sample handling more efficient.

The third dimension of separation involves off-line coupling of the 2D-LE device to nanocapillary reversed phase liquid chromatography (RPLC)[17,18] for a net 3D solution separation. Overall, this 3D solution separation platform can afford a separation peak capacity that rivals that of 2D gels, albeit with a ~2-4 fold higher loading requirement. The peak capacity from the 3D separation platform can be estimated for ubiquitin. The sIEF and GELFrEE devices each afford approximately five and 10 fractions, respectively. Since ubiquitin was detected in a single fraction from each dimension, the 2D-LE device affords a peak capacity of ~50. For RPLC, at baseline resolution, the peak capacity was about 50. With the assumption that each separation mode is perfectly orthogonal, a total peak capacity of 2,500 was estimated for the 3D separation platform for ubiquitin. In cases where separation resolution is unaffected, high loading should also translate to increases in the proteome coverage, although this was not rigorously evaluated here. Typical chromatographic resolution is displayed in **Fig. 1c (bottom right)**. In addition to online intact mass detection and fragmentation, the Fourier-Transform mass spectrometers (12 Tesla LTQ FT Ultra or the Orbitrap Elite, Thermo Fisher Scientific) can also be considered the fourth and final dimension of "separation" in the 4D platform. Of course, this resolution of protein isoforms occurs in the gas phase driven by either the ion trap or FTICR for separation based on mass/charge ratio. The "peak capacity" from the mass analyzer was estimated by assuming that each charge state occupies ~2 Th units (disregarding Fourier Transform MS (FTMS) ability to resolve isotopes). Since there are ~10 charge states with adequate intensity, this expands to 20 Th. In a 1,500 *m/z* window from 500-2,000, the ion trap or FTMS has a peak capacity of ~75 per MS^1 scan. Factoring the peak capacity from the 3D separation, and assuming each mode is perfectly orthogonal, the 4D resolving power has an approximate total peak capacity of 2,500 × 75 = 187,500. Images analogous to 2D gels were created using LC-MS-detection for the 2D-LE fractions and an example is shown in **Fig. 2a**. The software for this was recently described, and such heat maps were useful to compare the large data sets generated in this study[19].

Top Down Analytical Strategy To Provide High Proteome Coverage

Our overall goal in this study is first to provide large scale identifications of the human proteome with the highest content of molecular information preserved for the primary structures of endogenous proteins. This can be considered the discovery mode for interesting targets. High resolution FTMS was used for both precursor and fragmentation for fractions corresponding to MW <25 kDa enabling high quality protein characterization[17]. For increased sensitivity, the ion trap was used for detection of precursor ions above ~30 kDa, enabling protein isoform and species differentiation as described in detail elsewhere[18].

The total identification count (1043 unique proteins) in this study was accumulated from eight 4D separation runs (three nuclear and five cytosolic runs) and two GELFrEE-nanocapillary LC runs of samples enriched for mitochondrial membranes (**Supplementary Table 1**). From those identifications, 43% (447) were obtained in a single 4D analysis (**Supplementary Fig. 1a**) along with >1000 distinct isoforms fully or partially characterized in automated mode. This particular analysis of nuclear proteins from 5 to ~70 kDa (35 LC-MS/MS runs) was accomplished in 2 days and generated a total of 11,326 identification events with a ~44% spectral hit rate. From a single replicate analysis comprised of 4D runs of nuclear and cytosolic extracts, 611 unique protein identifications were obtained (**Supplementary Fig. 1b**) and these data were used to generate the quasi-2D gel[19] of **Fig. 2a**.

Protein Identification at High Mass

Using a data acquisition mode tailored for larger proteins, we were able to automatically identify proteins in the 40-110 kDa range (**Fig. 3d** and **Supplementary Table 2**). In fact, several examples were observed >60 kDa where the data contained intact mass values and bidirectional fragmentation patterns where both termini were localized unambiguously. This includes GRP 78, a 70.6 kDa heat shock protein with >12 fragment ions mapping to each terminus (**Supplementary Fig. 6**). The mass was determined to be 70550 Da, 25 Da lower than expected (70575.7 Da), assuming no modifications other than cleavage of its 18 amino acid signal peptide (**Supplementary Fig. 6**). Analyzing selected 2D-LE fractions containing 70-120 kDa proteins led to 10 further identifications in five LC-MS/MS runs (**Supplementary Table 2**). Together

with a report in 2006[4], this study shows the feasibility of high mass protein analysis in a proteomic setting. The main challenges at high mass presently are the separation of proteins >50 kDa by RPLC and the need for more intelligent data acquisition on current-generation mass spectrometers. It should be noted that the ability to exclude fragmentation of the same neutral mass during data dependent fragmentation (but with different charge states) was not utilized in this study. With the incremental advances projected by the "Moore's Law" of MS[20], the interrogation of 100 kDa isoforms will soon become comparable to measurements at 30 kDa today.

Targeted PTM Analysis Using Zoom Mapping and PTMCRAWLER

After obtaining a rigorous list of identifications, our next goal is to profile levels of changes in PTMs on proteins from cells undergoing the stress of DNA-damage. Our two-stage strategy was to analyze nuclear and cytosolic extracts in a mapping phase using 4D separation (striving for maximal proteome coverage), followed by targeted detection of interesting protein forms using only 3D separation and "zoom mapping" using the mass spectrometer (**Methods**). Since zoom mapping involves maximizing the FTMS cell with only the ions of interest, a dramatic S/N increase was observed versus broadband detection. To intelligently select intact masses for zoom mapping that are not part of the charge state distribution (*i.e.*, reselecting the same protein species), "mass mode" was enabled in the Xcalibur software. This ensured that each zoom map scan provided analysis on a different protein species.

After obtaining a rigorous list of intact masses from these experiments, our custom software, PTMCRAWLER was used to search for mass differences correlative to PTMs such as phosphorylations, methylations, and acetylations (**Fig. 2b**). The PTM profiling experiment involves re-running the same GELFrEE-RPLC fractions with the FT mode set for zoom mapping at a 60 *m/z* window (mass range is chosen to encompass the most intense charge state). Since RPLC has the ability to partially separate certain PTMs such as oxidations and phosphorylations, the time segment option in Xcalibur was used to ensure complete mass spectrometric profiling of a single protein before switching to a different target. Typically, 6-8 sets protein species are monitored for each LC-MS injection.

Isoelectric Focusing Enhances Phosphorylation Detection but Perturbs Quantitation

After using PTMCRAWLER software, our next step is to quantitate the levels of PTM targets. Our lab has shown that levels of PTMs can be semi-quantitative without the need for isotopic labelling[21]. The 4D separation did perturb phosphorylation stoichiometry since the IEF mode separates species by pI. This effect was shown on two targets in **Supplementary Fig. 9**. This prompted us to implement a simpler 3D separation platform where isoelectric focusing was omitted to maximize the likelihood that targeted protein species produced from the same gene are contained in the same GELFrEE fraction.

Ideally, a first-pass interrogation of intact proteins in a complex proteome would include detailed information on all protein species, including those containing previously uncharacterized mass shifts. A full description of this complexity requires multiple rounds of analysis and extensive tandem MS to map the "basis set" of expressed protein species (*cf.* section on cataloging protein species in the main text). In the full 4D approach, we observed partial fractionation of phosphoprotein species in the isoelectric focusing step due to small differences in pI values (*e.g.*, **Supplementary Fig. 9a** vs. **b**). In the case of BANF1 (accession number O75531), this effect enhanced the relative abundance of minor phosphorylated species, helping to determine the hierarchy of phosphorylations on multiply-modified phosphoproteins (**Supplementary Fig. 9d**). Once such protein species are known and characterized, three dimensions of separation (omitting IEF) was used to quantify species dynamics and assessing their function. Using just the 3D approach (no IEF), the data for HMGN1 (P05114), for example, more accurately portrayed the *in vivo* ratios of its phosphorylated forms by observing them all in the same mass spectrum (**Supplementary Fig. 9c**). The skewed intensity of the phosphorylation levels attributed to bias introduced by isoelectric focusing on the same protein can be compared (**Supplementary Fig. 9d**).

REFERENCES FOR SUPPLEMENTARY INFORMATION

1 Aston, F. W. Isotopes and atomic weights. *Nature* **105**, 617 (1920).

2 Futrell, J. H. Use of mass spectral data in radiation chemistry. *J. Chem. Phys.* **35**, 353-356 (1961).

3 Ruotolo, B. T. *et al.* Evidence for macromolecular protein rings in the absence of bulk water. *Science* **310**, 1658-1661 (2005).

4 Han, X., Jin, M., Breuker, K. & McLafferty, F. W. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **314**, 109-112 (2006).

5 Chait, B. T. Mass spectrometry: bottom-up or top-down? *Science* **314**, 65-66 (2006).

6 Jungblut, P., Holzhutter, H., Apweiler, R. & Schluter, H. The speciation of the proteome. *Chem. Cent. J.* 2:16 (2008).

7 Schluter, H., Apweiler, R., Holzhutter, H.G. & Jungblut, P. Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* 3:11 (2009).

8 Duncan, M. W., Aebersold, R. & Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28**, 659-664 (2010).

9 Boyne, M. T., Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Precise characterization of human histories in the H2A gene family by top down mass spectrometry. *J. Proteome Res.* **5**, 248-253 (2006).

10 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.* **5**, 233-239 (2006).

11 Zabrouskov, V. *et al.* Stepwise deamidation of ribonuclease A at five sites determined by top down mass spectrometry. *Biochemistry* **45**, 987-992 (2006).

12 Ge, Y., Rybakova, I. N., Xu, Q. G. & Moss, R. L. Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12658-12663 (2009).

13 Resemann, A. *et al.* Top-down de novo protein sequencing of a 13.6 kDa camelid single heavy chain antibody by matrix-assisted laser desorption ionization-time-of-flight/time-of-flight mass spectrometry. *Anal. Chem.* **82**, 3283-3292 (2010).

14 Tran, J. C. & Doucette, A. A. Rapid and effective focusing in a carrier ampholyte solution isoelectric focusing system: a proteome prefractionation tool. *J. Proteome Res.* **7**, 1761-1766 (2008).

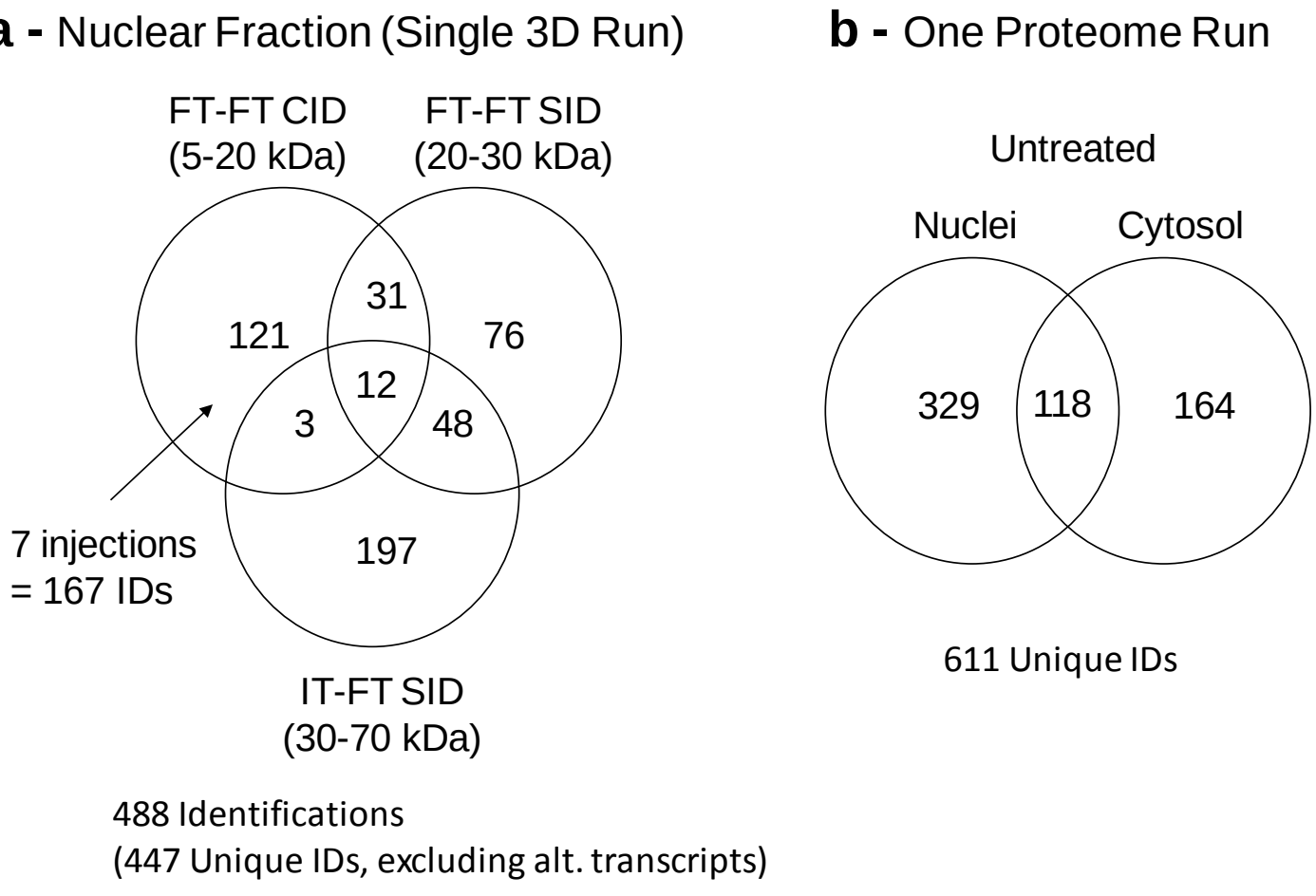
15 Tran, J. C. & Doucette, A. A. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* **80**, 1568-1573 (2008).

16 Tran, J. C. & Doucette, A. A. Multiplexed size separation of intact proteins in solution phase for mass spectrometry. *Anal. Chem.* **81**, 6201-6209 (2009).

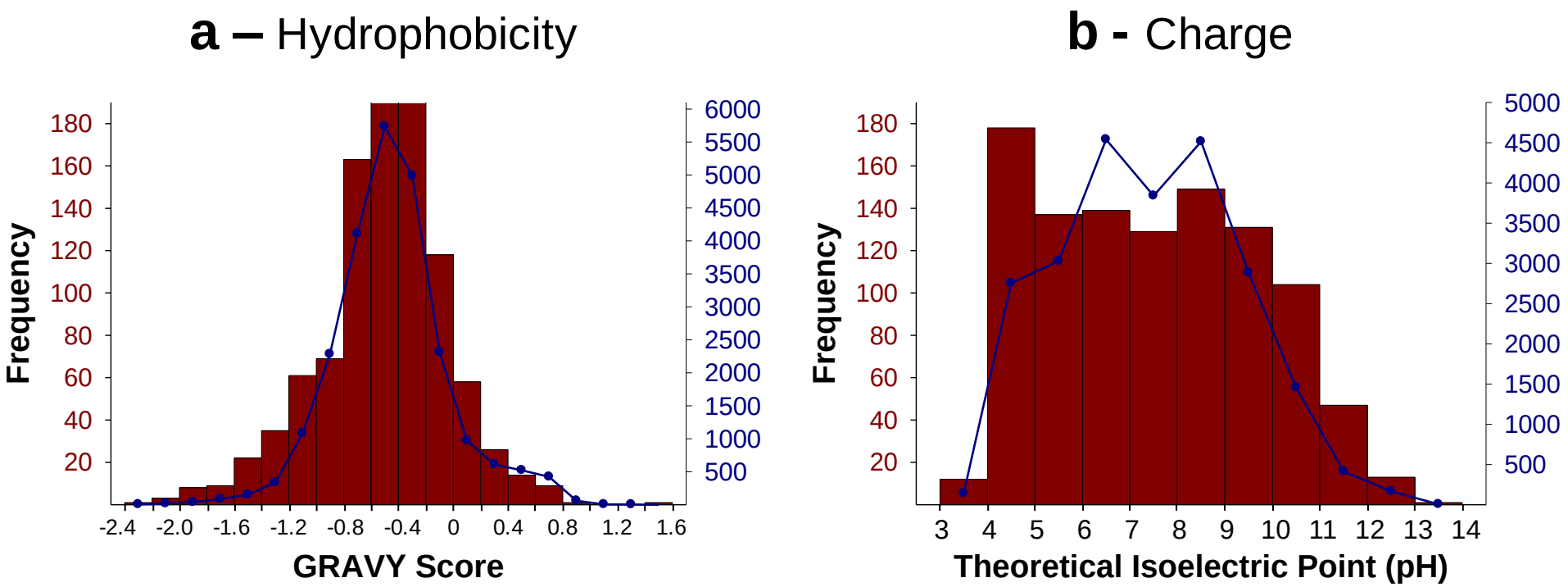
17 Lee, J. E. *et al.* A robust two-dimensional separation for top-down tandem mass spectrometry of the low-mass proteome. *J. Am. Soc. Mass Spectrom.* **20**, 2183-2191 (2009).

18 Vellaichamy, A. *et al.* Size-sorting combined with improved nanocapillary liquid chromatography-mass spectrometry for identification of intact proteins up to 80 kDa. *Anal. Chem.* **82**, 1234-1244 (2010).

19 Durbin, K. R. *et al.* Intact mass detection, interpretation, and visualization to automate top down proteomics on a large scale. *Proteomics **10***, 3589-3597 (2010).
20 The general analytical figures of merit in biological mass spectrometry (sensitivity, mass accuracy, and resolution) improve by about 3-4 fold every half decade or so. Michael W. Senko, *personal communication*.
21 Pesavento, J. J., A., M. C. & Kelleher, N. L. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone H4. *Anal. Chem.* **78**, 4271-4280 (2006).

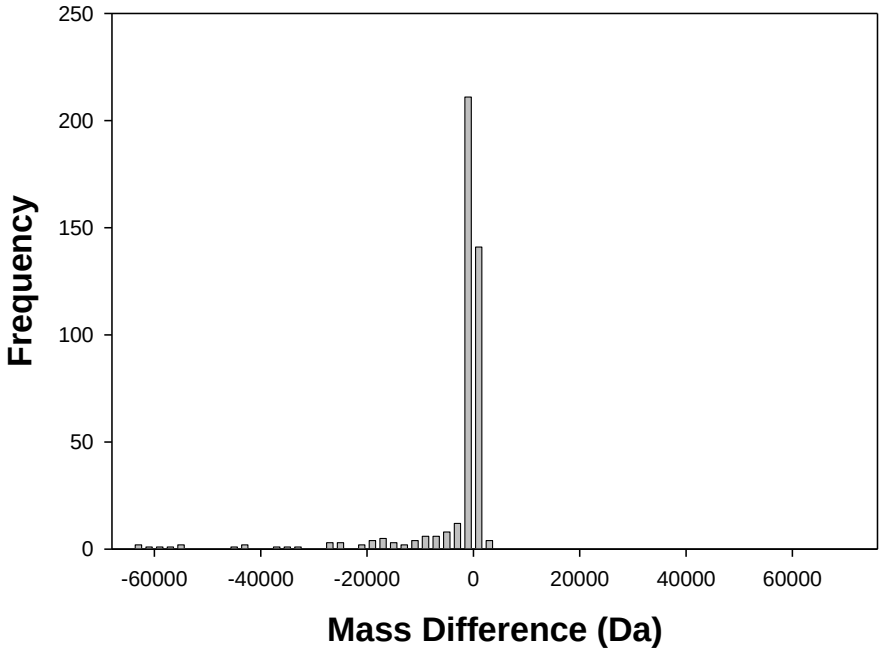


Supplementary Figure 1. (**a**) Venn diagram showing the number of proteins identified (including alternative transcripts) in each of the three modes of mass spectrometric data acquisition involving FT-FT-CID, FT-FT-SID and IT-FT-SID. Excluding alternative transcripts, 447 unique identifications were obtained in one single 4D run of nuclear extracts. (**b**) Venn diagram showing the overlapping number of unique proteins identified from a full proteome analysis which included both a nuclear and a cytosolic fraction subjected to a 4D run.

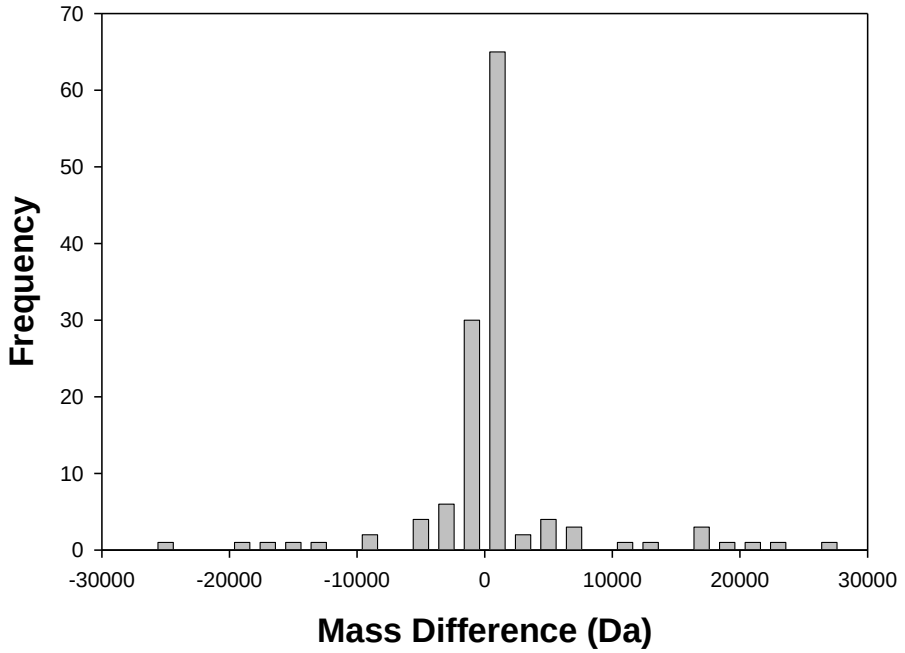


Supplementary Figure 2. Bar graphs showing the frequency of proteins identified in this study plotted against the theoretical (**a**) hydrophobicity and (**b**) isoelectric points. These plots are compared to the total theoretical distribution of the corresponding intrinsic property as predicted from the entire human proteome (blue line graph). The similarity of these distributions suggests that the proteins identified using the top down platform in this study have minimum bias against hydrophobicity and isoelectric point extremities.

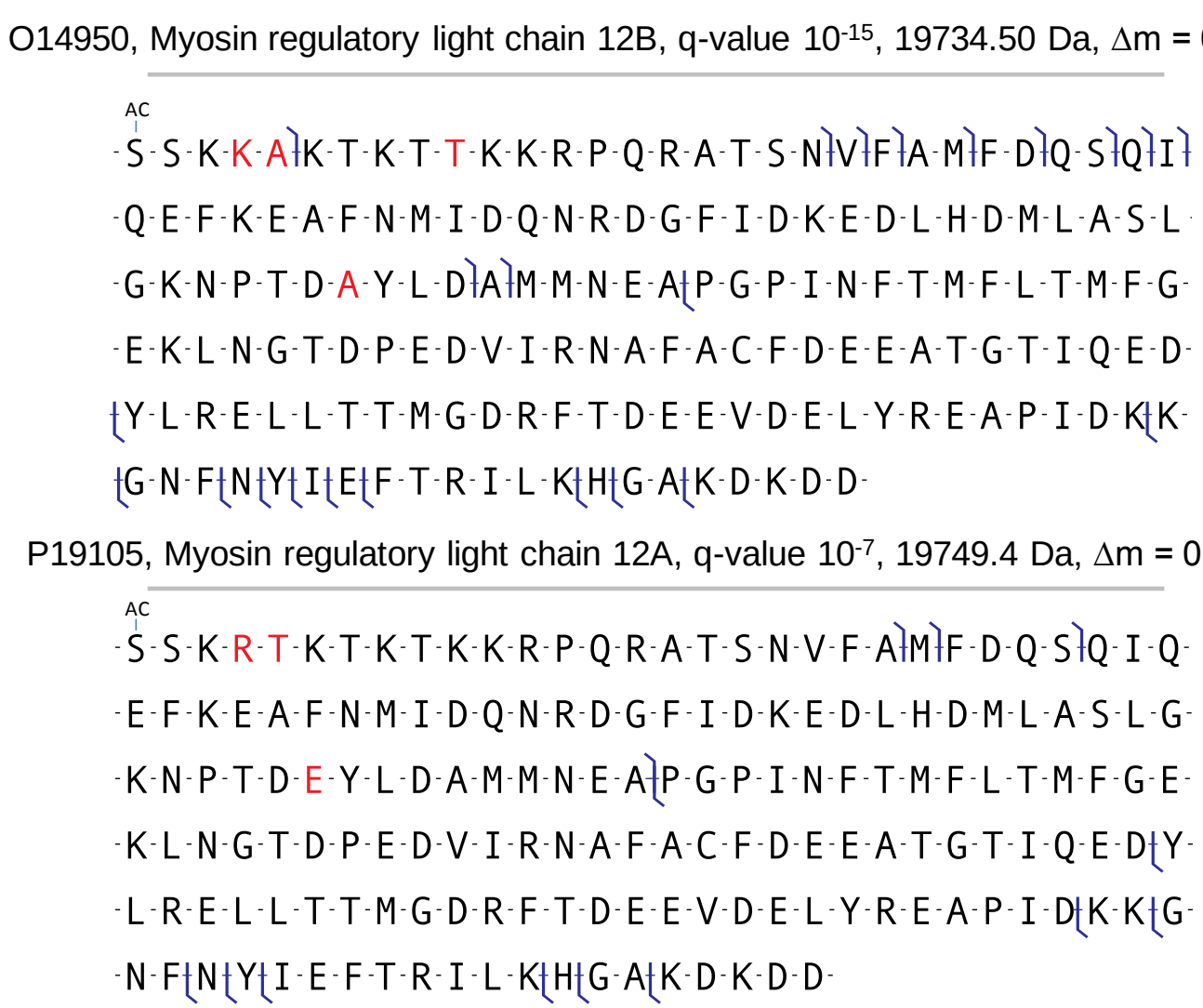
a – Deisotoped Data



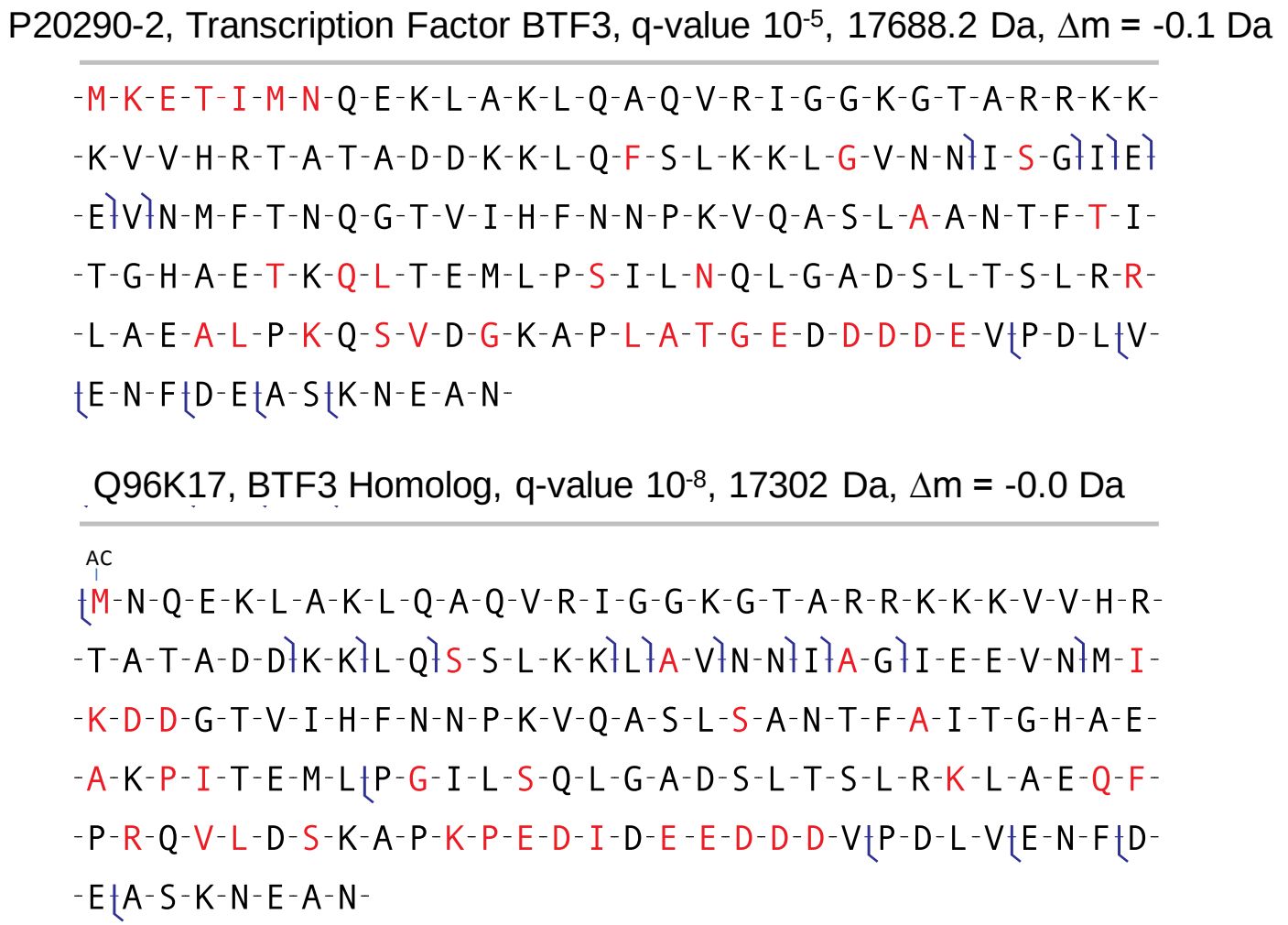
b – Deconvoluted Data



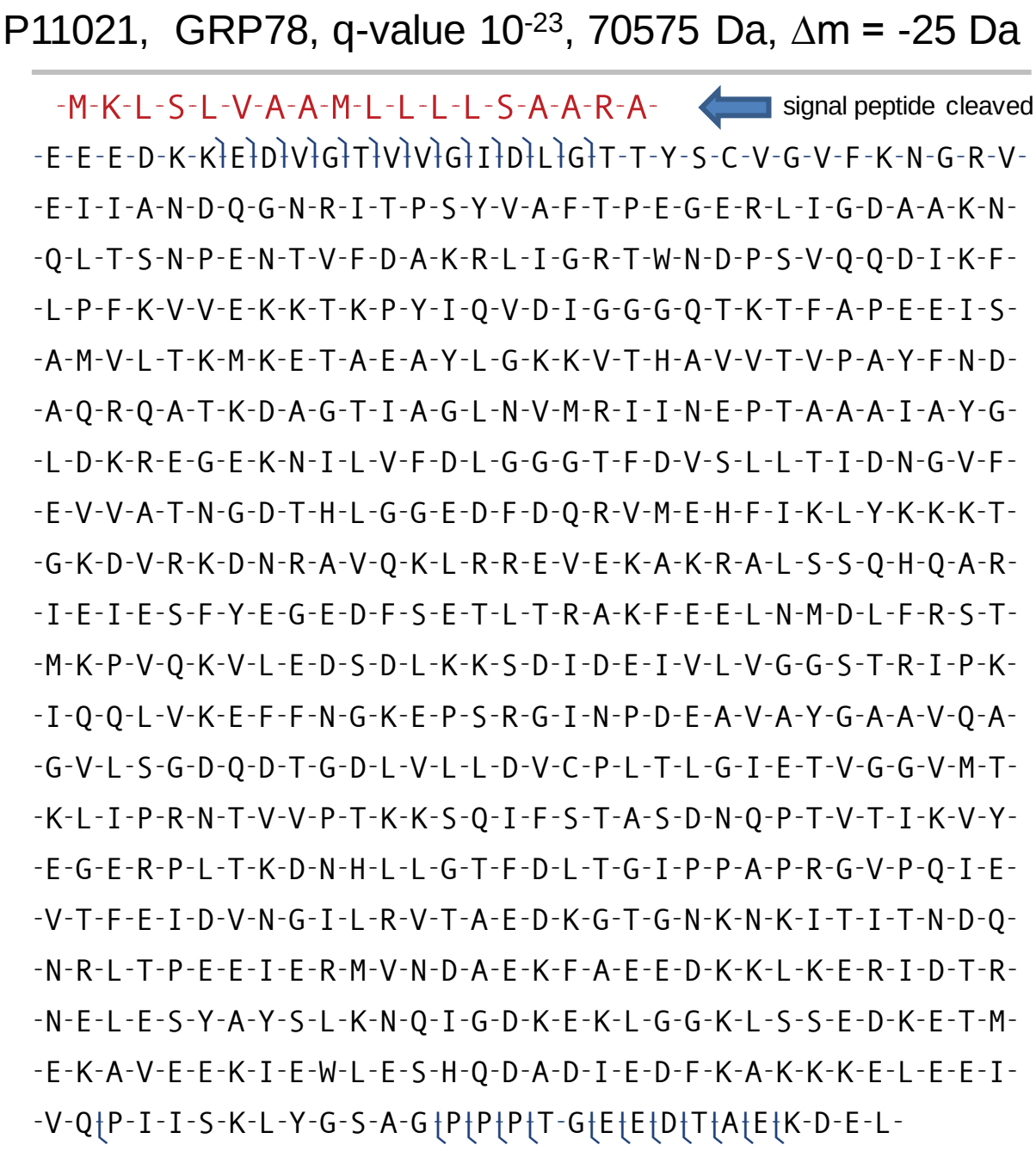
Supplementary Figure 3. Bar graphs of mass discrepancies showing the number of identifications with high quality intact mass values that deviated from the theoretical mass of the identified isoform. Graphs were obtained by plotting the (**a**) cases with isotopic resolution or (**b**) with masses determined by deconvolution of ESI charge states.



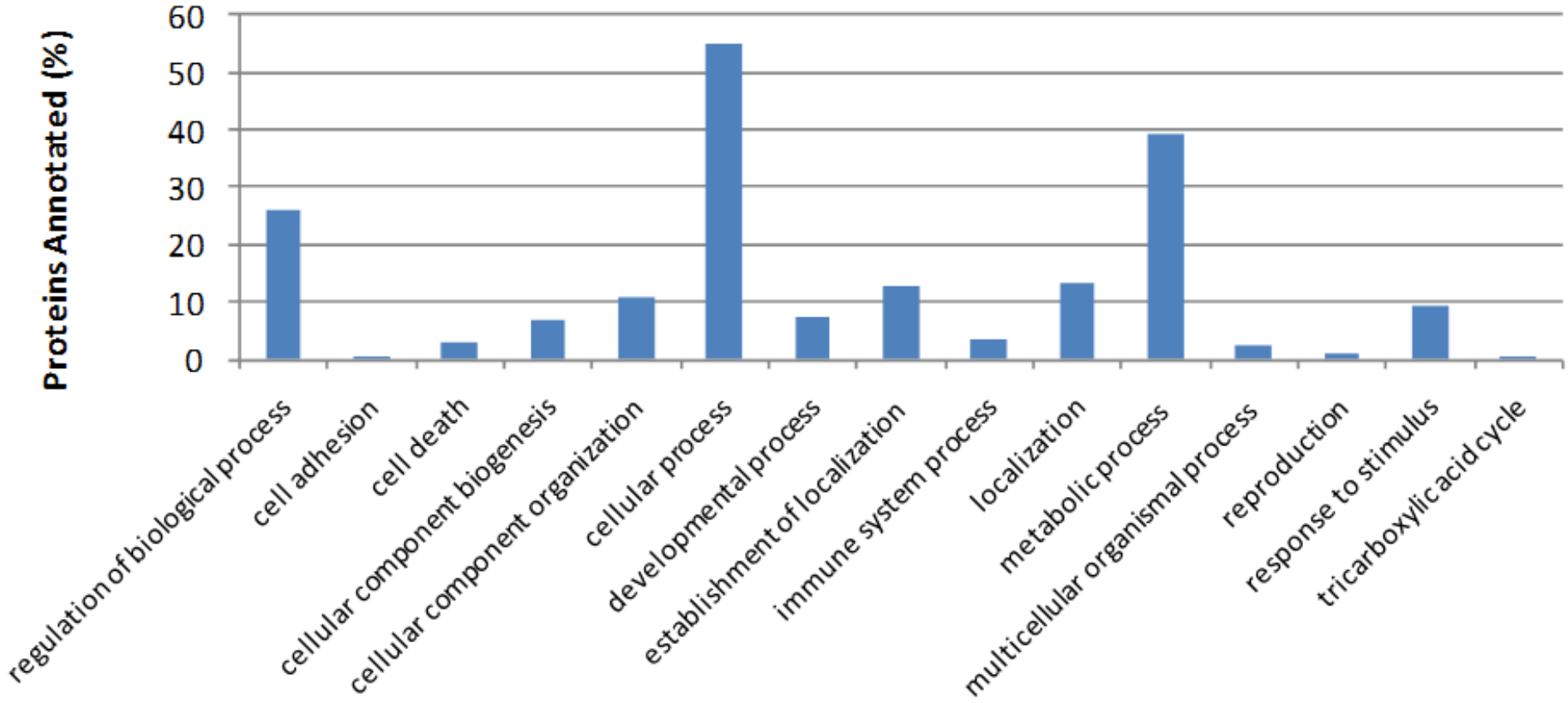
Supplementary Figure 4. Graphical fragmentation map of two myosin regulatory light chain proteins. Since the accurate intact masses of both these proteins were obtained, top down was able to differentiate these proteins despite the 97% sequence identity (differences highlighted in red).



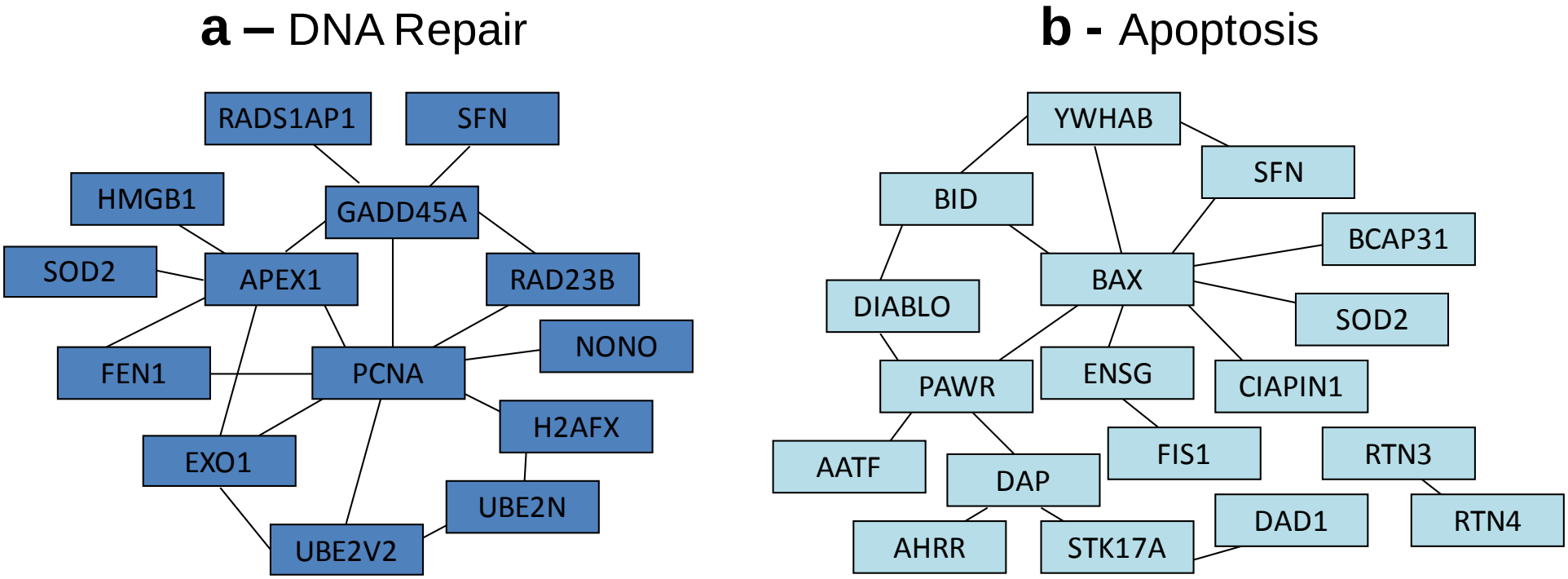
Supplementary Figure 5. Graphical fragmentation map of transcription factor BTF3 shows that top down MS/MS can differentiate between related proteins that arise from different human genes (differences highlighted in red).



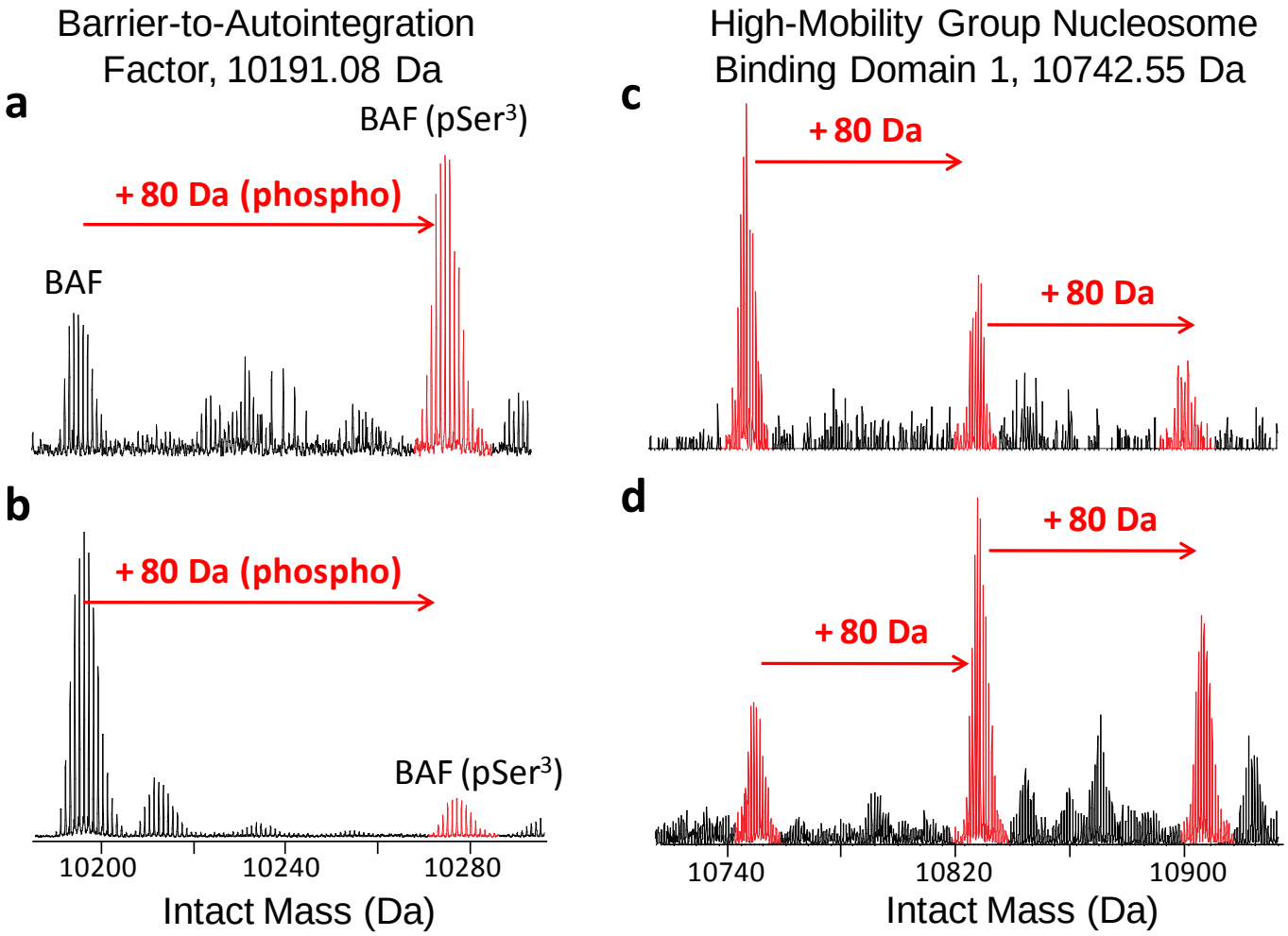
Supplementary Figure 6. Graphical fragmentation map of the protein GRP 78 reveals bidirectional fragmentation observed for this 71 kDa protein. The 18 amino acid sequence highlighted in red designates the signal peptide that was cleaved.



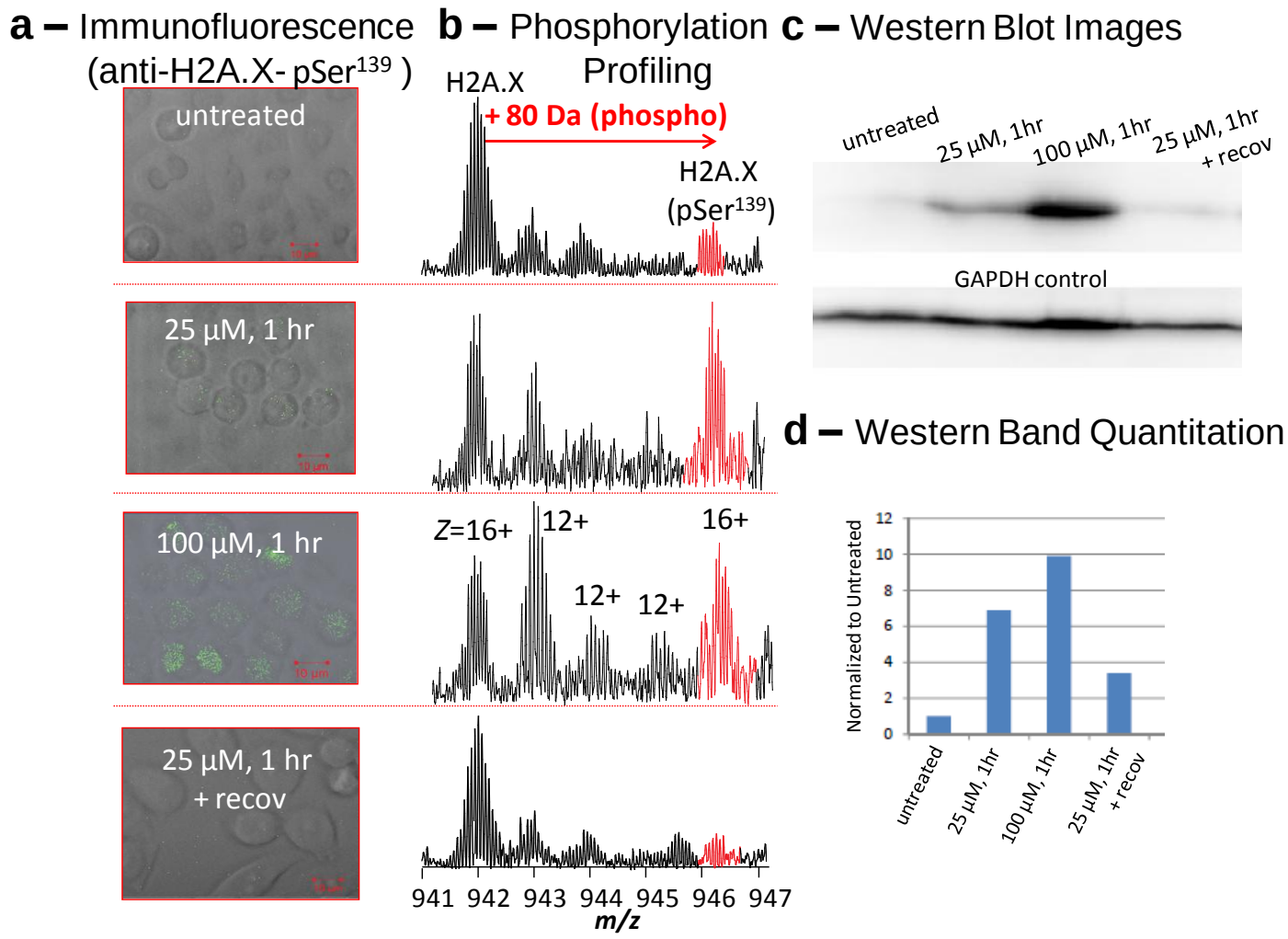
Supplementary Figure 7. Gene Ontology (GO) analysis of the proteins identified in this study showing the proportion of proteins that are annotated in the specified biological process.



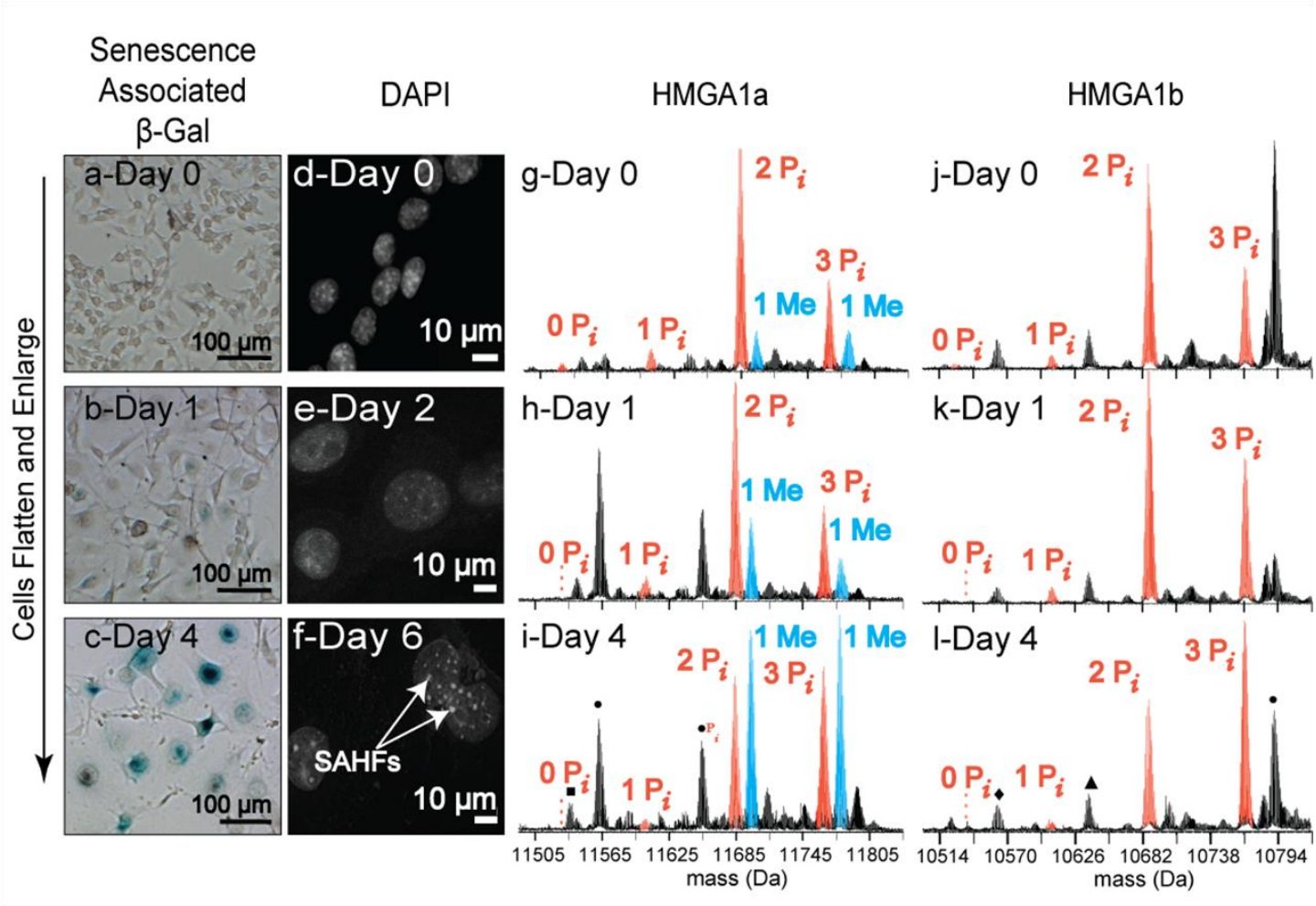
Supplementary Figure 8. Protein interaction networks using STRING analysis (Jensen, L. J. *et al. Nucleic Acids Res.* **37** D412-D416 (2009)) on the set of identifications obtained during this study for (**a**) DNA repair and (**b**) apoptosis.



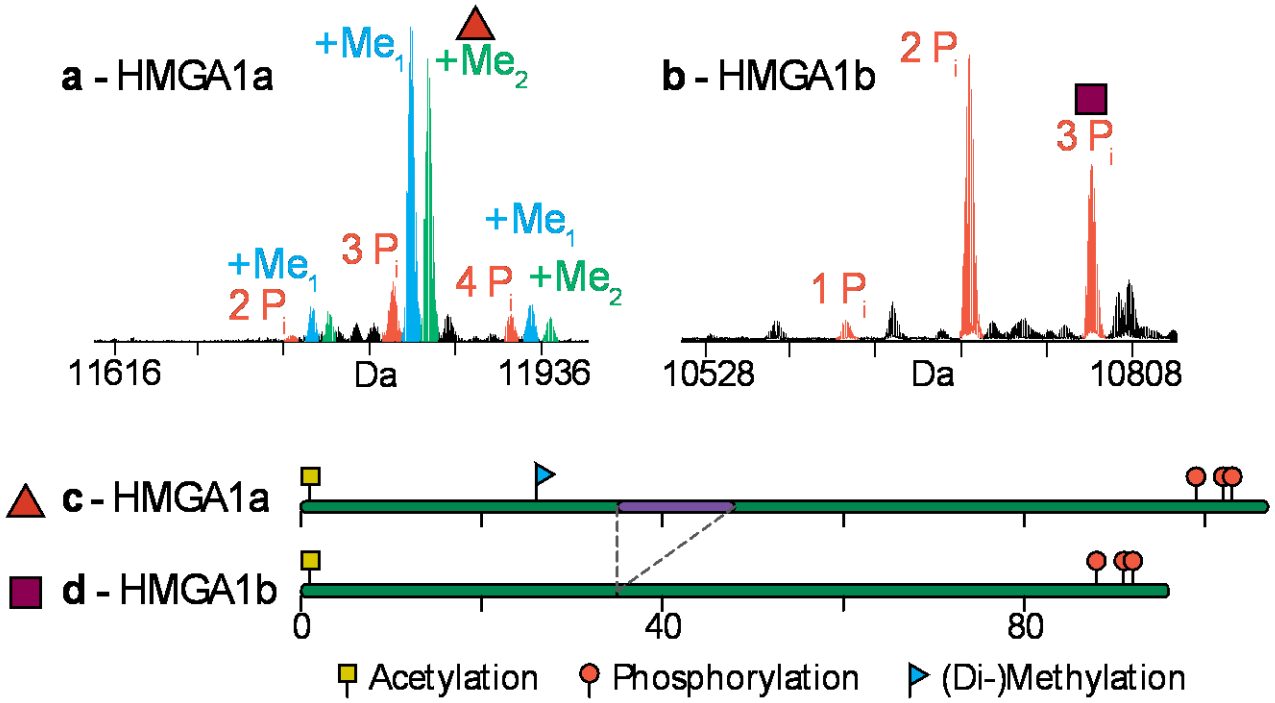
Supplementary Figure 9. MS profiles obtained from a 4D run depicting the ability of sIEF at separating phosphorylated BANF (O75531) from its unmodified form. As expected, the phosphorylated form is found in a more (**a**) acidic sIEF fraction (pH 3.66) versus the unmodified form, found in a more (**b**) basic fraction (pH 4.66). (**c**) To more accurately determine phosphorylation stoichiometry, sIEF was omitted, revealing a better measurement of the multiple modifications for HMGN1 (P05114) versus (**d**) the skewed phosphorylation levels as obtained when IEF was used.



Supplementary Figure 10. Phosphorylation profiling on γH2A.X during DNA damage. Confocal microscopy using anti-H2A.X-pSer139 (**a**) and mass spectrometric (**b**) profiling provides consistent results in agreement with both techniques. Both experiments reveal lower levels of γH2A.X in untreated HeLa cells versus DNA damaged cells (1 h with 25 µM or 100 µM etoposide). Repair of DNA damage (1 h, 25 µM etoposide followed by 24 h, recovery in fresh media) was observed with the decrease in phosphorylation levels (bottom panel). (**c**) These results were further confirmed with western blots monitoring the levels of γH2A.X. (**d**) Quantitation of the band intensities by normalizing the γH2A.X levels to the GAPDH was performed. In summary, MS profiling for γH2A.X shows ~50% increase in phosphorylation levels after DNA damage. In contrast, phosphorylation levels of only ~10% were observed in the cells that were untreated or had undergone DNA repair.



Supplementary Figure 11. Monitoring dynamics of related protein isoforms during senescence in B16F10 melanoma cells. (**a-c**) After induction of DNA damage by transient treatment with the topoisomerase II inhibitor, etoposide, X-Gal staining over a four day recovery period revealed robust activity of β-galactosidase (a known marker of senescence). (**d-f**) Chromatin compaction and nuclear expansion shown by staining DNA with DAPI. After several days of recovery, senescence associated heterochromatic foci (SAHFs) are clearly observed as punctate regions of fluorescence. Changes in modification profiles on HMGA1a (**Panels g-i**) and HMGA1b **(Panels j-l**) splice variants over the period of four days after etoposide treatment. HMGA1a shows marked increase in the relative abundance of mono-methylated isoform while HMGA1b remains unmethylated. The neutral mass labels on the x axis correspond only to proteins with the same charge state as the HMGA1A form. Unrelated protein masses in spectra: ♦, 7544 Da; ●, 10637 Da; ▲, 10020 Da; ■, 10767 Da.



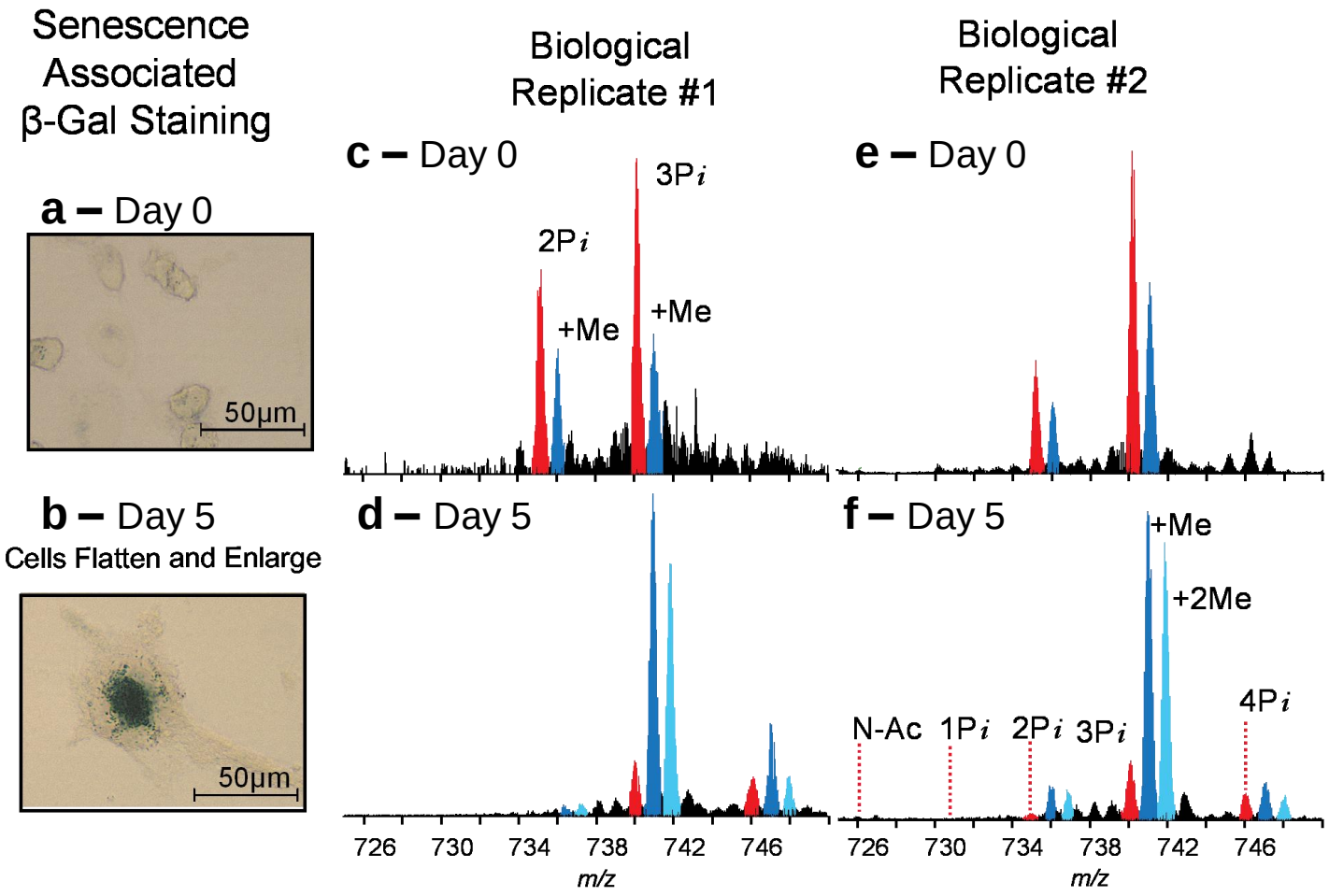
e - HMGA1a Graphical Fragment Map

-S-E-S-G-S-K-S-S-Q-P-L-A-S-K-Q-E-K-D-G-T-E-K-R-G-R-G-R-P-R-K-
-Q-P-P-V-S-P-G-T-A-L-V-G-S-Q-K-E-P-S-E-V-P-T-P-K-R-P-R-G-R-P-
-K-G-S-K-N-K-G-A-A-K-T-R-K-V-T-T-A-P-G-R-K-P-R-G-R-P-K-K-L-E-
-K-E-E-E-E-G-I-S-Q-E-S-S-E-E-E-Q-

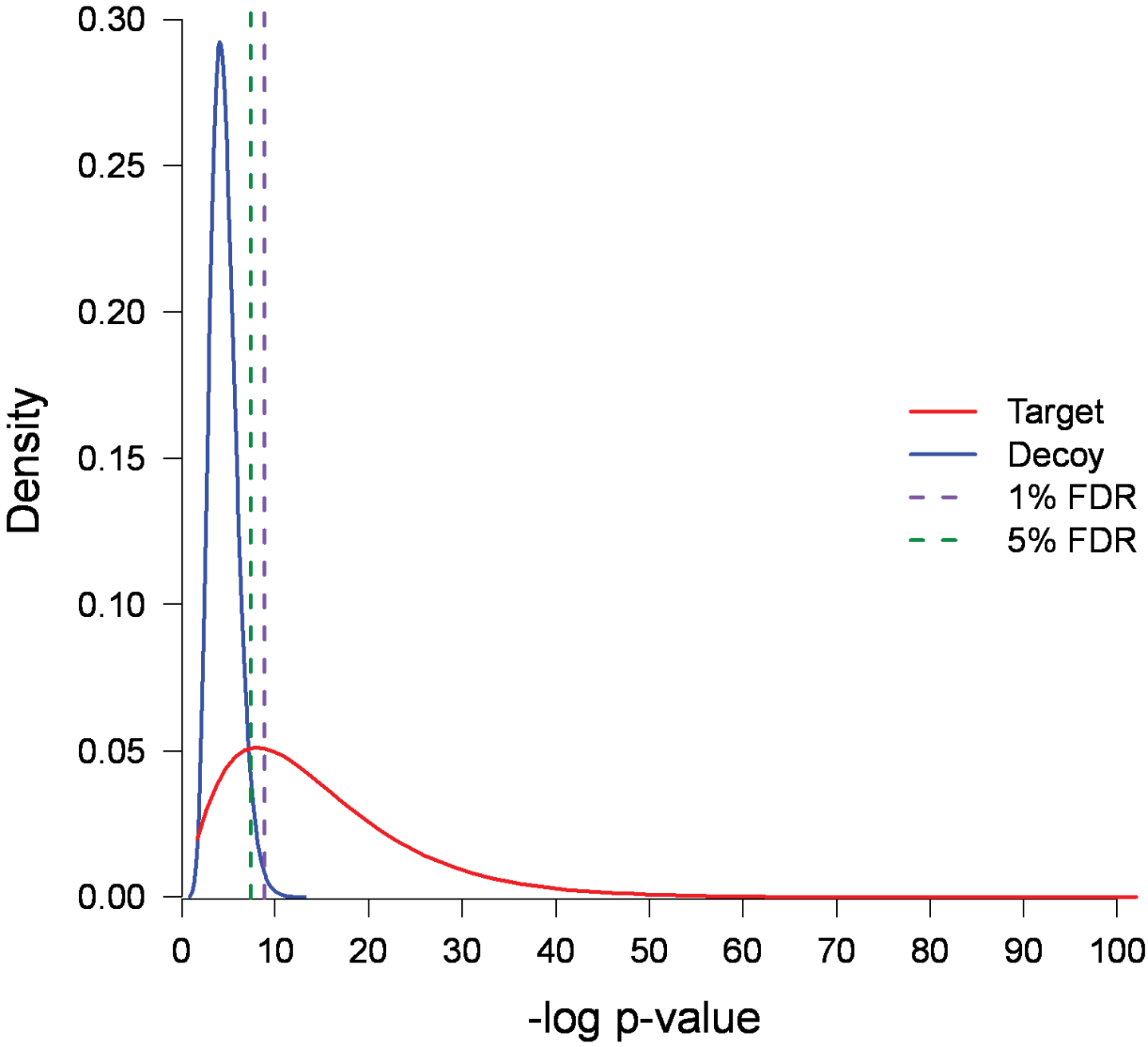
f - HMGA1b Graphical Fragment Map

-S-E-S-G-S-K-S-S-Q-P-L-A-S-K-Q-E-K-D-G-T-E-K-R-G-R-G-R-P-R-K-
-Q-P-P-K-E-P-S-E-V-P-T-P-K-R-P-R-G-R-P-K-G-S-K-N-K-G-A-A-K-T-
-R-K-V-T-T-A-P-G-R-K-P-R-G-R-P-K-K-L-E-K-E-E-E-E-G-I-S-Q-E-S-
-S-E-E-E-Q-

Supplementary Figure 12. Characterization of two HMGA1 isoforms (indicated with a triangle, ▲, and a square ■). Isoform profiles are shown for (**a**) HMGA1a and (**b**) HMGA1b. Diagram highlighting the sites of alternative splicing and post-translational modification present on the (**c**) HMGA1a (tri-phosphorylated, di-methylated form) and (**d**) HMGA1b (tri-phosphorylated form). (**e**) Graphical fragment map showing precise localization of the di-methylation site at Arg25 and the addition of 11 amino acids present in HMGA1a but absent in (**f**) HMGA1b.



Supplementary Figure 13. Phosphorylation and methylation levels on HMGA1a isoforms increase with senescence. (**a**) Light microscopy of H1299 cells in culture. (**b**) After induction of DNA damage on H1299 cells by transient treatment with camptothecin, β-Gal staining over a 5 day recovery period revealed robust activity of β-galactosidase (a known marker of senescence). (**c-d**) MS profiles of the HMGA1a showed significant increases in phosphorylation in addition to a large increase in both mono- and di-methylation levels during senescence. (**e-f**) A biological replicate experiment showed reproducible results.



Supplementary Figure 14. Probability score distributions used in the calculation of *q-values*. The distributions of both the decoy (n = 133,426) and target (n = 95,483) data are plotted and modelled with gamma functions. The Poisson-based p-scores (corresponding to the *q-values* used for identification thresholds at 1% and 5% FDRs) are shown with dotted lines.