LETTER

doi:10.1038/nature10575

Mapping intact protein isoforms in discovery mode using top-down proteomics

John C. Tran[1,2], Leonid Zamdborg[1], Dorothy R. Ahlf[1,2], Ji Eun Lee[1,3], Adam D. Catherman[1,2], Kenneth R. Durbin[1,2], Jeremiah D. Tipton[2], Adaikkalam Vellaichamy[1]†, John F. Kellie[1,2], Mingxi Li[1,2], Cong Wu[1], Steve M. M. Sweet[1,2], Bryan P. Early[1,2], Nertila Siuti[1]†, Richard D. LeDuc[4], Philip D. Compton[2], Paul M. Thomas[1,2] & Neil L. Kelleher[1,2]

A full description of the human proteome relies on the challenging task of detecting mature and changing forms of protein molecules in the body. Large-scale proteome analysis[1] has routinely involved digesting intact proteins followed by inferred protein identification using mass spectrometry[2]. This 'bottom-up' process affords a high number of identifications (not always unique to a single gene). However, complications arise from incomplete or ambiguous[2] characterization of alternative splice forms, diverse modifications (for example, acetylation and methylation) and endogenous protein cleavages, especially when combinations of these create complex patterns of intact protein isoforms and species[3]. 'Top-down' interrogation of whole proteins can overcome these problems for individual proteins[4,5], but has not been achieved on a proteome scale owing to the lack of intact protein fractionation methods that are well integrated with tandem mass spectrometry. Here we show, using a new four-dimensional separation system, identification of 1,043 gene products from human cells that are dispersed into more than 3,000 protein species created by post-translational modification (PTM), RNA splicing and proteolysis. The overall system produced greater than 20-fold increases in both separation power and proteome coverage, enabling the identification of proteins up to 105 kDa and those with up to 11 transmembrane helices. Many previously undetected isoforms of endogenous human proteins were mapped, including changes in multiply modified species in response to accelerated cellular ageing (senescence) induced by DNA damage. Integrated with the latest version of the Swiss-Prot database[6], the data provide precise correlations to individual genes and proof-of-concept for large-scale interrogation of whole protein molecules. The technology promises to improve the link between proteomics data and complex phenotypes in basic biology and disease research[7].

Effective fractionation[8–10] is critical for sample handling before mass-spectrometry-based proteomics. So far, no fractionation procedure for intact proteins can match the resolution of two-dimensional gel electrophoresis (two-dimensional gels). Here we use a liquid phase alternative to two-dimensional gels that bypasses both their low recovery and extensive workup steps before mass spectrometry[11]. This procedure for two-dimensional liquid electrophoresis[12] comprises solution isoelectric focusing (sIEF) followed by gel-eluted liquid fraction entrapment electrophoresis (GELFrEE)[13] for fractionation by protein isoelectric point and size, respectively (Fig. 1a, b). Combining these with nanocapillary liquid chromatography and mass spectrometry (LC–MS) (Fig. 1c) for both low[14] and high molecular mass proteins[15] results in an overall four-dimensional separation of whole protein molecules before ion fragmentation by tandem mass spectrometry and protein identification.

Using the four-dimensional platform described above, we generated a quasi-two-dimensional gel perspective of the human proteome with extremely high molecular detail (Fig. 2a) from individual replicate

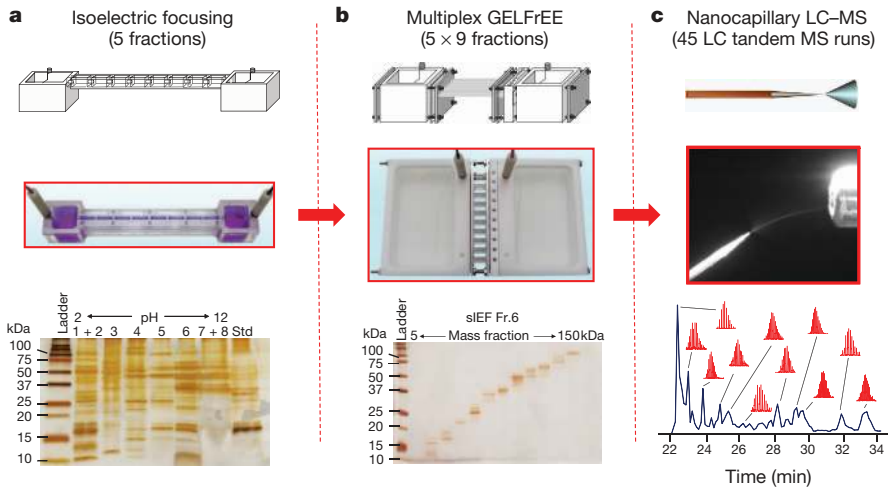

Figure 1 | The four-dimensional platform for high-resolution fractionation of protein molecules. Schematics (top) and photographs (middle) are shown for (**a**) a custom device for sIEF, (**b**) a custom device for multiplexed GELFrEE and (**c**) RPLC coupled to mass spectrometry. Representative one-dimensional gels of fractions collected from the two electrophoretic devices are shown below their pictures; note the resolution attainable at the level of intact proteins. The combined resolution of RPLC with Fourier-transform mass spectrometry is depicted by the chromatogram along with selected isotopic distributions for protein ions measured during the run.

[1]Departments of Chemistry and Biochemistry, and the Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [2]Departments of Chemistry and Molecular Biosciences and the Feinberg School of Medicine, Northwestern University, Evanston, Illinois 60208, USA. [3]Doping Control Center and Center for Theragnosis, Korea Institute of Science and Technology, Seoul, South Korea. [4]Department of Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. †Present addresses: Department of Nanomedicine, The Methodist Hospital Research Institute, Houston, Texas 77030, USA (A.V.); Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA (N.S.).

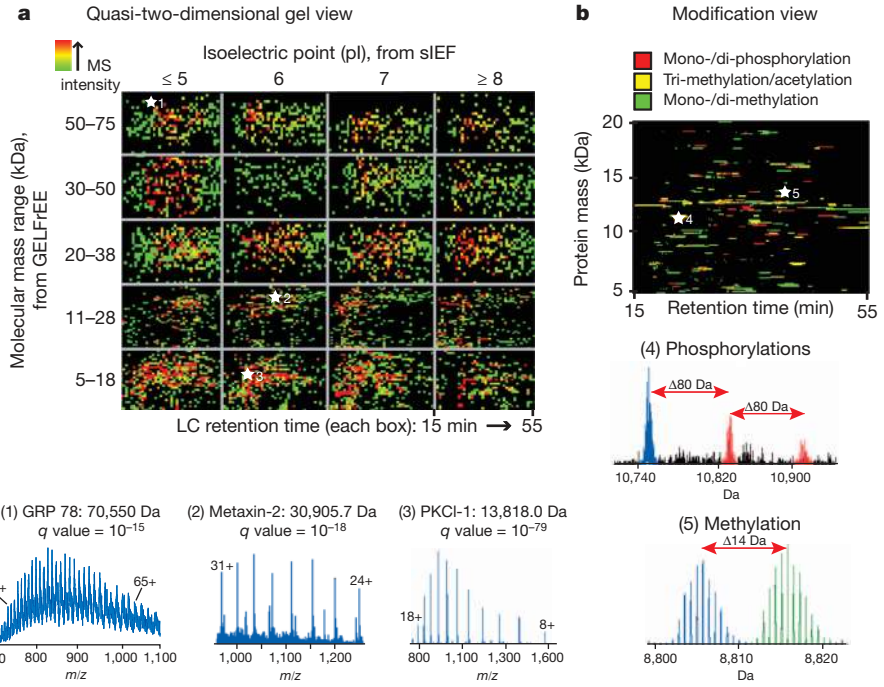

Figure 2 | Two visual representations of proteome-scale runs. **a**, The heat map is generated from combined four-dimensional runs of nuclear and cytosolic extracts. Intact mass and isoelectric point values are indicated on the *y*- and *x*-axes, respectively. Each box in the grid displays total ion chromatograms from LC runs of two-dimensional liquid electrophoresis fractions plotted as time versus neutral intact mass. The mass spectrometry intensity is indicated by colour (legend on top left). Representative precursor scans were extracted from the heat map for electrospray ionization–mass spectrometry spectra of high (1), medium (2) and low (3) mass proteins, along with their identifications from online fragmentation (insets at bottom). **b**, Plot created from selective display of protein pairs with mass differences consistent with acetylation (yellow), phosphorylation (red) and methylation (green), with three and two protein species shown as examples in insets (4) and (5), respectively.

analyses of nuclear and cytosolic extracts of HeLa S3 cells (Supplementary Fig. 1). In discovery mode, the IEF–GELFrEE-nanocapillary liquid chromatography platform used 0.5–1 mg of input protein and provided a peak capacity of well over 2,000 for separation of protein molecules in solution. Considering the separation power of the mass spectrometer, the peak capacity of the four-dimensional system is greater than 100,000 for proteins below approximately 25 kDa (Supplementary Information). This is 20-fold higher than the peak capacity for high-resolution two-dimensional gels (less than 5,000). Identification and characterization of isoforms were achieved using fragmentation data acquired with less than 10 part-per-million mass accuracy for searching databases with highly annotated primary sequences[16]. Using tailored software[17], we overcame the 'protein inference problem' where identification ambiguity results when isoforms (for example, from members of a gene family or alternative splicing) produce many identical tryptic peptides[2,18]. The databases and search engine used here are fully compatible with the UniProt flat file format and enable a deep consideration of known PTMs, alternative splice variants, polymorphisms, endogenous proteolysis and diverse combinations of all these sources of molecular variation at the protein level[16]. Together with the careful curation of the Swiss-Prot database[6], the result is an informatics framework that maps each given protein identification to a single gene (except in rare cases like ubiquitin, where multiple genes can produce the identical sequence). Extended details on statistical analysis are provided in the Methods section.

A total of 1,043 proteins were identified with unique Swiss-Prot accession numbers in this study (Supplementary Table 1). These identifications originate from 1,045 human genes, 77% of whose protein products displayed amino (N)-terminal acetylation. The distribution of *q* values, which indicates the confidence of protein identifications (see Methods), is shown in Fig. 3c. This level of proteome coverage represents the most comprehensive implementation of top-down mass spectrometry so far, with an approximate tenfold increase in identifications of intact proteins for any microbial system[19–21] and a greater than 20-fold increase over any previous work in mammalian cells[14,22] (Fig. 3a). In addition, fragmentation evidence for 3,093 protein isoforms/species was captured in this initial report (Supplementary Table 1), with PTMs detected as follows: 645 phosphorylations, 538 lysine acetylations, 158 methylations, 19 lipid/terpenes and 5 hypusines. Over 400 species were attributed to core histones alone. Comparisons of predicted protein hydrophobicity and isoelectric point showed minimal bias compared with that expected for the human proteome (Supplementary Fig. 2).

Using an orthogonal method to detect PTMs based on intact mass values[17], we detected pairs of protein species showing characteristic mass differences (Fig. 2b). For proteins less than 20 kDa, 225 pairs showed mass differences consistent within 0.05 Da with mono-methylation, 185 with di-methylation and 122 with tri-methylation/acetylation. Other mass differences revealed 87 cases consistent with double acetylation, 140 with mono-phosphorylation and 100 with di-phosphorylation events (Fig. 2b). Using this set of mass differences on the entire HeLa data set for all isotopically resolved proteins, a total of 2,130 such mass shifts were found.

Complete characterization of a protein requires the theoretical and experimental mass values to match within error. For the 1,043 proteins identified, 431 and 331 were identified with intact mass information from either isotope spacings or deconvolution of charge states, respectively. Of these data, 54% of the isotopically resolved proteins matched the species identified from the database within 2 Da (Supplementary Fig. 3a). Likewise, 130 of 331 of the masses determined by deconvolution were manually determined to be of high quality and 51% of these matched within 200 Da (Supplementary Fig. 3b). The protein species outside these windows are clearly identified by fragmentation, but harbour unexplained mass discrepancies (Δm) at this time. The complete explanation of Δm in the human proteome motivates future refinements in data acquisition to obtain enough tandem mass spectrometry information on all the protein isoforms/species.

Major functional differences can exist among protein isoforms in a family, making their precise identification a major boost in the

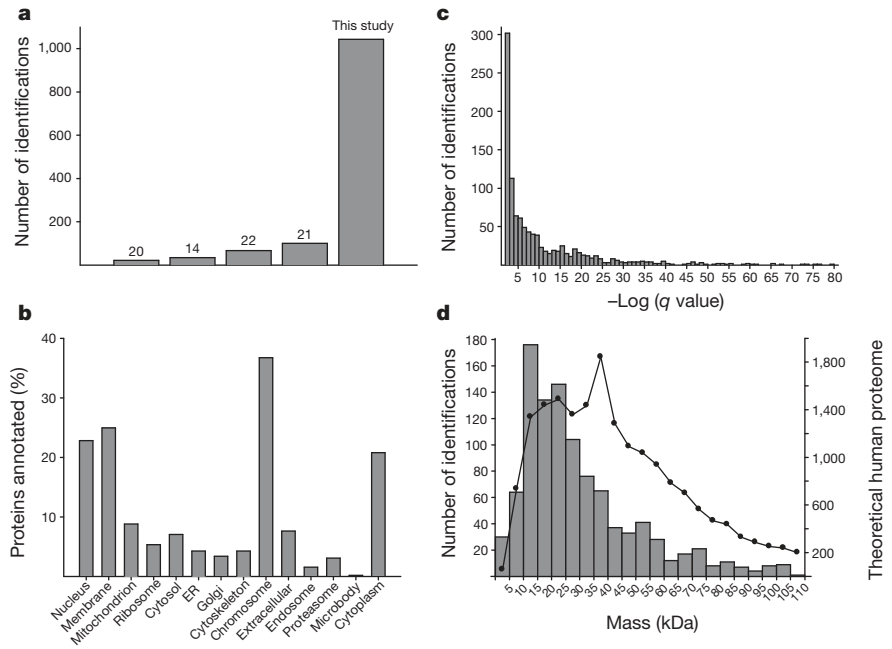

Figure 3 | Proteome analysis metrics associated with this study. **a**, Graph showing the striking increase in identifications from previous studies achieved in archaeal, bacterial, yeast or human systems. References cited are indicated above each bar. **b**, A gene ontology analysis for the identifications in this study. **c**, Histogram showing the distribution of *q* values for the identified proteins. **d**, Plot showing the molecular mass distribution for the unique identifications obtained. The line graph depicts the theoretical molecular mass distribution for the human proteome (Swiss-Prot, *Homo sapiens*, 20,223 entries).

information content of proteomic analyses in higher eukaryotes. An intact protein mass and matching fragment ions from both termini are usually sufficient to accomplish a gene-specific identification[4,17]. Here, 9 of the approximately 15 isoforms of histone H2A were fully characterized in an automated fashion despite their greater than 95% sequence identity (including the H2A.Z and H2A.X variants) with an additional three having Δm greater than 1 Da (H2A type 1-D, 2-C and 2-B). Also identified were nine S100 proteins, several α- and β-tubulins, seven unique isoforms of human keratin (a widely known contaminant in proteomics), MLC20, BTF3 and their related sequences (which are 97% and 81% identical, respectively (Supplementary Figs 4 and 5)), and over 100 isoforms/species from the high mobility group (HMG) family (see, for example, Fig. 4). Significant improvements for top-down proteomics in discovery mode were made for proteins in the 40–110 kDa range (Fig. 3d), including extensive characterization of GRP78, a 70.6-kDa heat-shock protein (more than 12 fragment ions mapping to each terminus, Supplementary Fig. 6), and identification of several proteins greater than 90 kDa, such as P33991 and Q14697 at 97 and 104 kDa, respectively (Supplementary Table 2).

Because the two-dimensional liquid electrophoresis platform makes use of SDS extensively, we anticipated reduced bias against integral membrane proteins. In all, 32% of the 1,043 total identifications from HeLa cells were membrane-associated proteins (GO:0016020), with 62% of these annotated as integral membrane proteins (GO:0016021, Supplementary Table 4). A more focused study of a mitochondrial membrane fraction (see Methods) used chromatographic procedures modified for enhanced separation of membrane proteins. We identified an additional 46 integral membrane proteins (Supplementary Table 3) from a single three-dimensional experiment (no isoelectric focusing). Detailed inspection of the species that eluted from the column during LC–MS revealed proteins with a distribution of 1–11 transmembrane helices (Supplementary Table 3). This shows a broad applicability of this study and will drive further efforts to detect full-length isoforms of membrane proteins[23].

As part of our study of the HeLa proteome, cells were treated with etoposide to elicit the DNA damage response (see Methods), followed by four-dimensional fractionation and top-down tandem mass spectrometry. Using gene ontology (GO) analysis, we annotated all

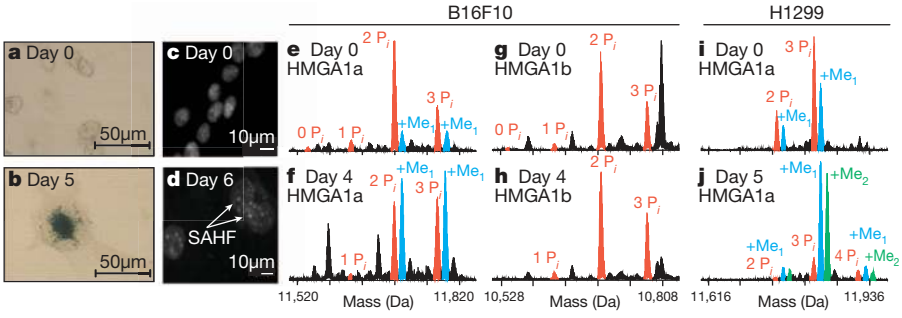

Figure 4 | Monitoring dynamics of HMGA1 isoforms during senescence in B16F10 and H1299 cells. After induction of DNA damage by transient treatment with camptothecin for H1299 cells or etoposide for B16F10, progression of accelerated senescence was monitored by SA-β-Gal (**a**, **b**) or DAPI staining to monitor formation of senescence-associated heterochromatic foci (SAHF) (**c**, **d**) over the specified recovery period. Changes in modification profiles on HMGA1a (**e**, **f**) and HMGA1b (**g**, **h**) from B16F10 showed mild increases in phosphorylation occupancy but a significant increase in methylation levels on multiply phosphorylated species. A more striking increase in both methylation and phosphorylation was observed in senescent H1299 cells (**i**, **j**). No such methylations were observed in the HMGA1b profiles for either cell line.

four-dimensional identifications according to cell compartment (Fig. 3b) or biological process (Supplementary Fig. 7). Many proteins detected were involved in cell cycle regulation and apoptosis, including nine that interact with proliferating cell nuclear antigen during repair of DNA damage (Supplementary Fig. 8). Also, several proteins involved in the Fanconi anaemia pathway were identified including FANCE, RAD51AP1, RAD23B and RPA3, with the last two completely characterized (Supplementary Table 5). Several CDK inhibitors were found, such as $p27^{Kip1}$ (CDKN1B) and $p16^{INK4a}$ (CDKN2A), T53G1 and the protein product from a target gene of p53 (Q9Y2A0, p53-activated protein 1).

Using the three-dimensional fractionation approach (that is, GELFrEE-nanocapillary LC–MS) to readout phosphorylation stoichiometry with high fidelity (Supplementary Information and Supplementary Fig. 9), we monitored 17 phosphoprotein targets across three time points at three different concentrations of etoposide (Supplementary Table 6). We found increases in the occupancy of phosphorylation in H2A.X-pSer139 (γH2A.X) after treatment with 25 or 100 μM etoposide for 1 h (Supplementary Fig. 10). After a 24-h recovery from treatment, a return to basal levels of phosphorylation of γH2A.X was found, consistent with engagement of the DNA-repair machinery[24]. Further, we observed a strong correlation between the phosphorylation stoichiometry of γH2A.X determined by mass spectrometry with the results from immunofluorescence and western blotting run in parallel (Supplementary Fig. 10a–c).

In separate studies we tracked over 2,300 species (from 690 proteins) in H1299 cells (Supplementary Table 7) and 2,300 species (from 708 proteins) in B16F10 melanoma cells (Supplementary Table 8) in the days after a 24-h treatment with camptothecin or 5 h of etoposide, respectively, using only the three-dimensional fractionation approach. After induction of DNA damage, we also monitored the classic hallmarks of stress-induced senescence in H1299 (ref. 25) and B16F10 (ref. 26) over several days (Supplementary Fig. 11a–c), including cell enlargement and formation of senescence-associated heterochromatic foci (Supplementary Fig. 11d–f). Although levels of γH2A.X remained the same as in control cells, a striking upregulation in methylated forms of di- and tri-phosphorylated HMGA1a, but not of its splice variant HMGA1b was observed as both B16F10 and H1299 cells entered stress-induced senescence (Fig. 4 and Supplementary Fig. 11g–l).

Full descriptions of the fragmentation data for two multiply modified species of HMGA1 are presented in Supplementary Fig. 12. In mapping these species, the hierarchy of phosphorylations on HMGA1a was determined for control cells to be Ser 101 and Ser 102 occupied in the $2P_i$ form and evidence for the third site pointing predominantly towards pSer 98. The $3P_i$ and $4P_i$ forms both showed some occupancy for pSer 43 (data not shown), a site only available in the splice region specific to the HMGA1a variant (Supplementary Fig. 12). For day 5 in senescent H1299 cells, the effect on methylation was particularly dramatic, with both the mono- and di-methylated species (also harbouring multiple phosphorylations) reproducibly increased to be greater than 80% of the total signal for species from the *hmga1* gene (Fig. 4 and see Supplementary Fig. 13 for biological replicates). The methylation site was localized precisely to Arg 25 (Supplementary Fig. 12), consistent with previous work on HMGA1 proteins[27]. A similar response for methylated HMGA species has been observed in damaged cancer cells undergoing apoptosis[27,28] but the B16F10 and H1299 cells prepared here were clearly senescent as measured by annexin V staining and fluorescence-activated cell sorting analysis through day 6 (data not shown). As Arg 25 is in the first AT-hook DNA-binding region (residues 21–31), it is possible that the R25me1 and R25me2 marks perturb DNA-kinking and allow HMGA1a to be preferentially incorporated into senescence-associated heterochromatic foci[29] during accelerated cellular senescence. Other changes in bulk chromatin were also notable, such as hypoacetylation on all core histones, increased levels of H3.2K27me2/3, and decreased H3.2K36me3.

The sharp increase in proteome coverage demonstrated here provides a path ahead for interrogating the natural complexity of protein primary structures that exist within human cells and tissues. Because this is the first time top-down proteomics has been achieved at this scale, an early glimpse at the prevalence of uncharacterized mass shifting events has been revealed in the human proteome. With faithful mapping of intact isoforms on a proteomic scale, detecting co-variance in modification patterns will help lay bare the post-translational logic of intracellular signalling. Also, proper speciation of protein molecules offers the promise of increased efficiency for biomarker discovery through stronger correlations between measurements and organismal phenotype (for example, a particular isoform of apolipoprotin C-III and levels of high-density lipoprotein in human blood[7]). Technology for intact protein characterization could also become a central approach to focus an analogous effort to the human genome project—to provide a definitive description of protein molecules present in the human body[30].

METHODS SUMMARY

For large-scale global analysis, HeLa S3 cells were pre-fractionated using custom two-dimensional liquid electrophoresis platform, comprising sIEF coupled to multiplexed GELFrEE[12,13]. HeLa S3, H1299, B16F10 cells and mitochondrial membrane proteins were also fractionated using the custom GELFrEE[13] device alone (no sIEF). After separation, detergent and salt were removed, and the fractions were injected into nanocapillary reversed-phase liquid chromatography (RPLC) columns for elution into a 12 T linear ion trap Fourier-transform mass spectrometer for online detection and fragmentation[14,15]. The mass spectrometry RAW files were processed with in-house software called CRAWLER to assign masses. Using this program, determination of both the intact masses and the corresponding fragment masses was performed and these data were searched against a human proteome database. Extensive statistical workups were also performed using several false discovery rate (FDR) estimation approaches (with decoy databases both concatenated and not). A final q value procedure is described in detail (Methods), with the data above reported using a 5% instantaneous FDR (that is, q value) cutoff at the protein level (Supplementary Fig. 14).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 March; accepted 15 September 2011.
Published online 30 October; corrected 7 December 2011 (see full-text HTML version for details).

1. Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6,** 359–362 (2009).
2. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440 (2005).
3. Schluter, H., Apweiler, R., Holzhutter, H. G. & Jungblut, P. R. Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* **3,** 11 (2009).
4. 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).
5. 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. USA* **106,** 12658–12663 (2009).
6. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature Methods* **7,** 681–685 (2010).
7. Mazur, M. T. *et al.* Quantitative analysis of intact apolipoproteins in human HDL by top-down differential mass spectrometry. *Proc. Natl Acad. Sci. USA* **107,** 7728–7733 (2010).
8. Righetti, P. G., Castagna, A., Antonioli, P. & Boschetti, E. Prefractionation techniques in proteome analysis: the mining tools of the third millennium. *Electrophoresis* **26,** 297–319 (2005).
9. Wang, H. *et al.* Intact-protein-based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol. Cell. Proteomics* **4,** 618–625 (2005).
10. Capriotti, A. L., Cavaliere, C., Foglia, P., Samperi, R. & Laganà, A. Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. *J. Chromatogr. A* (in the press).
11. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA* **97,** 9390–9395 (2000).
12. Tran, J. C. & Doucette, A. A. Multiplexed size separation of intact proteins in solution phase for mass spectrometry. *Anal. Chem.* **81,** 6201–6209 (2009).
13. 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).

14. 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).
15. 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).
16. Roth, M. J. *et al.* Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol. Cell. Proteomics* **4,** 1002–1008 (2005).
17. 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).
18. Duncan, M. W., Aebersold, R. & Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nature Biotechnol.* **28,** 659–664 (2010).
19. Bunger, M. K., Cargile, B. J., Ngunjiri, A., Bundy, J. L. & Stephenson, J. L. J. Automated proteomics of *E. coli* via top-down electron-transfer dissociation mass spectrometry. *Anal. Chem.* **80,** 1459–1467 (2008).
20. Parks, B. A. *et al.* Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers. *Anal. Chem.* **79,** 7984–7991 (2007).
21. Patrie, S. M. *et al.* Top down mass spectrometry of 60-kDa proteins from *Methanosarcina acetivorans* using quadrupole FTMS with automated octopole collisionally activated dissociation. *Mol. Cell. Proteomics* **5,** 14–25 (2006).
22. Roth, M. J., Parks, B. A., Ferguson, J. T., Boyne, M. T. I. & Kelleher, N. L. 'Proteotyping': Population proteomics of human leukocytes using top down mass spectrometry. *Anal. Chem.* **80,** 2857–2866 (2008).
23. Gomez, S. M., Nishio, J. N., Faull, K. F. & Whitelegge, J. P. The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Mol. Cell. Proteomics* **1,** 46–59 (2002).
24. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316,** 1160–1166 (2007).
25. Roberson, R. S., Kussick, S. J., Vallieres, E., Chen, S. Y. & Wu, D. Y. Escape from therapy-induced accelerated cellular senescence in p53-null lung cancer cells and in human lung cancers. *Cancer Res.* **65,** 2795–2803 (2005).
26. Yawata, T. *et al.* Identification of a $\leq$ 600-kb region on human chromosome 1q42.3 inducing cellular senescence. *Oncogene* **22,** 281–290 (2003).
27. Sgarra, R. *et al.* During apoptosis of tumor cells HMGA1a protein undergoes methylation: identification of the modification site by mass spectrometry. *Biochemistry* **42,** 3575–3585 (2003).
28. Sgarra, R. *et al.* The AT-hook of the chromatin architectural transcription factor high mobility group A1a is arginine-methylated by protein arginine methyltransferase 6. *J. Biol. Chem.* **281,** 3764–3772 (2006).
29. Narita, M. *et al.* A novel role for high-mobility group A proteins in cellular senescence and heterochromatin formation. *Cell* **126,** 503–514 (2006).
30. Service, R. F. Proteomics ponders prime time. *Science* **321,** 1758–1761 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank all members of the group who contributed to development of top-down mass spectrometry over the years along with several private foundations: The Searle Scholars Program, The Burroughs Wellcome Fund, The David and Lucile Packard Foundation, The Richard and Camille Dreyfus Foundation, and The Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust. We further acknowledge the Department of Chemistry at the University of Illinois, the Neuroproteomics Center on Cell to Cell Signaling supported through the National Institute on Drug Abuse (P30DA 018310), the National Institute for General Medical Sciences (GM 067193-08) and the National Science Foundation (DMS 0800631), whose combined investment in basic research over the past decade made this work possible. We dedicate this work in memory of Jonathan Widom.

Author Contributions Project design: J.C.T., L.Z., P.M.T., N.L.K. Cell culture and biology: J.C.T., J.E.L., A.D.C., D.R.A., M.L., C.W., S.M.M.S., N.S. Separations: J.C.T., J.E.L., A.D.C., D.R.A. Mass spectrometry: J.C.T., J.E.L., A.D.C., D.R.A., J.D.T., A.V., J.F.K., P.D.C. Data analysis and statistics: J.C.T., L.Z., K.R.D., B.P.E., R.D.L., P.M.T., N.L.K. Writing: J.C.T., N.L.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.L.K. (n-kelleher@northwestern.edu).

METHODS

Cell culturing and treatments. HeLa S3 cells (ATCC CCL-2.2) were grown in Joklik's modified minimal essential medium. HeLa cells were grown in suspension, whereas B16F10 mouse melanoma cells (ATCC CRL-6475) and H1299 small lung carcinoma cells (ATCC CRL-5803, in DMEM) were grown adherently. Media were supplemented with 10% newborn calf serum and 1% penicillin and streptomycin. Cells were maintained in 5% CO_2 at 37 °C, harvested (after adding 0.05% trypsin and EDTA for B16F10 and H1299 cells) by centrifugation at 200*g* for 5 min, and washed twice with PBS.

For experiments using HeLa with intrinsic DNA damage, cells were treated for 1 or 5 h using 1 μM etoposide (four-dimensional platform), or for 1 h with 25 or 100 μM etoposide for targeted experiments (three-dimensional platform, Supplementary Fig. 10 and Supplementary Table 6)[31]. Repair of DNA damage was monitored after placing treated (25 μM) cells into fresh media for 24 h before harvesting. B16F10 cells were treated with 10 μM etoposide for 5 h at 20–30% confluence, and allowed to grow in normal media. Similarly, H1299 cells were treated with 25 nM camptothecin for 24 h. Over several days, these treatments induced stress-associated, accelerated senescence as monitored by[32] a flattened and enlarged cell morphology, expression of senescence-associated β-galactosidase (SA-β-gal), formation of senescence-associated heterochromatic foci and upregulation of p53 (in B16F10 cells only). Approximately 2×10^7 senescent cells were collected, lysed and subjected to the three-dimensional platform.

Preparation of HeLa S3 cytosolic, nuclear and whole-cell extracts. For large-scale mapping using the four-dimensional platform, the HeLa cytosolic and nuclear extracts were prepared through a protocol by Trinkle-Mulcahy *et al.*[33]. After isolation, the pelleted fraction containing nuclei was redissolved using 4% SDS (50 mM Tris, pH 7.5, with protease, phosphatase inhibitors and sodium butyrate). Both cytosolic and nuclear fractions were disrupted using a sonication probe. The nuclear fractions were further homogenized using QIAShredder homogenizer spin columns to reduce viscosity (Qiagen). All fractions were centrifuged at 14,000*g* for 10 min at 4 °C. Protein concentrations were determined by bicinchoninic acid and stored at −80 °C.

For the three-dimensional platform, whole-cell extracts were re-suspended in 5 ml of lysis buffer (4% SDS, 100 mM Tris-HCl pH 7.5, 10 mM DTT with protease, phosphatase inhibitors and sodium butyrate). The mixture was vortexed for 5 min and boiled for 10 min. Immediately after boiling, the samples were alkylated in the dark with 100 mM iodoacetamide for 20 min.

Western blots, imaging and microscopy. Western Blots. Antibodies used were Histone H2A.X-pSer139 (Cell Signaling Technology; 2577S), GAPDH (Santa Cruz; sc-47724) and HRP-conjugated secondary antibodies. Chemiluminescence was detected using ChemiDoc XRS+ (Bio-Rad Laboratories) and band densities were calculated using Image Lab software (Bio-Rad Laboratories).

For γH2A.X imaging by immunofluorescence, cells were fixed in 0.1% glutaraldehyde and 3% formaldehyde made fresh from paraformaldehyde followed by permeabilization in 0.5% Triton X-100. Cells were subsequently incubated for 1 h at 25 °C with 3% BSA and then with the primary antibody against Histone H2A.X-pSer139 (1:400). Alexa Fluor Conjugates (Molecular Probes) were used as the secondary antibodies (1:1,000). Images were obtained using an inverted Zeiss Axio Observer.Z1 confocal microscope.

For β-Gal staining, cells were fixed for 5 min in 2% formaldehyde/0.2% glutaraldehyde in PBS, washed and incubated at 37 °C (no CO_2) overnight with X-Gal (5-bromo-4-chloro-3-indolyl β-D-glactosidase) staining solution (1 mg of X-Gal in 40 mM citric acid/sodium phosphate pH 6.0, 5 mM potassium ferrocyanide, 5 mM potassium ferricyanide, 150 mM NaCl, 2 mM $MgCl_2$).

For DAPI staining, cells were fixed in 4% paraformaldehyde for 30 min, washed and incubated with 4′,6′-diamidino-2-phenylindole (DAPI) solution (10 μg ml^{-1}) for 10 min. The images of DAPI-stained DNA were obtained using the Zeiss Axio Observer.Z1 confocal microscope.

Sample handling and multidimensional protein fractionation. Fractionation using four-dimensional fractionation (two-dimensional liquid electrophoresis and LC–MS). HeLa proteins (0.5–2 mg) were reduced, alkylated, precipitated with cold acetone and re-suspended in 3.2 ml sIEF buffer (8 M urea, 2 M thiourea, 50 mM DTT, 1% w/v Biolyte 3/10 carrier ampholytes from Bio-Rad Laboratories). The sample was focused using a custom designed eight-chamber sIEF system as previously described[34]. After complete focusing (approximately 1.5 h at 2 W), the liquid fractions (400 μl) were collected and combined with their respective chamber rinse solution (100 μl of 1% SDS). Adjacent sIEF fractions (including anode and cathode) were pooled, resulting in about five fractions that were precipitated using cold acetone. The precipitated proteins were re-suspended in approximately 50 μl of Laemmli loading buffer[35]. These fractions were then fractionated in parallel using a custom ten-channel multiplexed GELFrEE device[12]. Tube gels were cast to 12% T (1 cm length) for the resolving and 4% T for the stacking gels (300 μl volume). Application of 240 V for approximately 1 h resulted in eight or nine GELFrEE fractions (150 μl) per IEF fraction after elution of the dye front. After complete electroelution, the two-dimensional liquid electrophoresis fractions underwent SDS removal using chloroform/methanol/water precipitation as described previously[36]. Before nanocapillary RPLC injection, fractions were re-suspended by pipetting vigorously with 15–40 μl solvent A (5% acetonitrile, 0.2% formic acid).

For experiments using the three-dimensional platform, only GELFrEE coupled to nanocapillary LC–MS was used for sample fractionation. Whole cell lysates, mitochondrial membrane preparations or extracts targeting modified proteins were re-suspended as described above and fractionated using a single-channel GELFrEE device[13]. Nanocapillary LC–MS conditions for large-scale analyses were as described below.

For nanocapillary RPLC–MS, in either the four- or three-dimensional platform, re-suspended fractions were injected (10 μl) onto a trap column (150 μm × 2 cm) using an autosampler (Eksigent). The nanobore analytical column (75 μm × 10 cm) containing an integral fritted nanospray emitter (PicoFrit, New Objective) was coupled to the trap in a vented column tee setup. Both the analytical and trap columns contain polymeric reversed-phase (PLRP-S, Phenomenex) media (5 μm, 1,000 Å pore size). The Eksigent 1D Plus nano-HPLC system was operated at a flow rate of approximately 2 μl min^{-1} for 10 min for loading onto the trap. The proteins were eluted into the mass spectrometer using a flow rate of 300 nl min^{-1} with the following gradient: 5% B (95% acetonitrile + 0.2% formic acid) at 0 min; 20% B at 5 min; 55% B at 50 min; 85% B at 55 min; 5% B at 65 min; 5% B at 75 min.

For proteins fractionated from HeLa S3 and B16F10 cells, the nanocapillary RPLC column was coupled online to a 12 T LTQ FT Ultra mass spectrometer (Thermo Fisher Scientific) fitted with a digitally controlled nanospray ionization source (PicoView DPV-550, New Objective). For masses up to 25 kDa (as determined from GELFrEE fractions), precursor mass data were collected using the Fourier-transform ion cyclotron resonance (eight microscans, 170,000 resolving power at $m/z = 400$) with an m/z range of 500–1,800 and a target value of 1 million charges. For masses either greater than 25 kDa or greater than 50 kDa, precursor mass data were obtained using the ion trap at 20 or 50 microscans, respectively. Data from H1299 cells were obtained on an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) using two to four microscans, 108,000 or 216,000 resolving power at $m/z = 400$, and a target value of 1 million charges.

Mass spectrometry data acquisition for targeted monitoring of intact isoforms/species. For three-dimensional experiments on targets up to 25 kDa, precursor mass data were collected with Fourier-transform ion cyclotron resonance parameters as described above. Data-dependent 'zoom mapping' was performed using a top three (no fragmentation) acquisition strategy with 60 m/z isolation window, three microscans at 85,000 resolving power (target value of 2 million charges with 60 m/z isolation at zero collision energy, or SIM mode). Dynamic exclusion was enabled with a repeat count of 2, an exclusion duration of 5,000 s, and a repeat duration of 240 s. 'Mass mode' was enabled in the Xcalibur software to ensure that each zoom map scan detected a different protein species.

Preparation of mitochondrial membrane proteins for top-down mass spectrometry. A HeLa S3 cell pellet consisting of approximately 10^9 cells was re-suspended in about 16 ml STM buffer (250 mM sucrose, 50 mM Tris-HCl, pH 7.4, 5 mM $MgCl_2$, 10 mM sodium butyrate, 1 mM DTT, 1% protease and phosphatase inhibitors (Sigma-Aldrich)). Cells were lysed using a glass Dounce homogenizer. The lysate was centrifuged at 800*g* for 15 min to remove nuclei and cellular debris. Mitochondria membrane isolation was performed as described previously[37]. Briefly, the cell lysate was centrifuged at 6,000*g* for 15 min. The pellet was washed with STM buffer and the centrifugation step repeated. Mitochondria were re-suspended in 2 ml lysis buffer (10 mM HEPES, pH 7.9, 10 mM sodium butyrate, 1 mM DTT, 1% protease and phosphatase inhibitors) and stirred at 4 °C for 30 min before sonication. The suspension was centrifuged at 9,000*g* for 30 min. The pellet was re-suspended in 0.5 ml extraction buffer (20 mM Tris, pH 7.8, 0.4 M NaCl, 15% glycerol, 5% SDS, 10 mM sodium butyrate, 1 mM DTT, 1% protease and phosphatase inhibitors) and vortexed for 30 min. The sample was centrifuged at 9,000*g* for 30 min and the supernatant was collected, aliquoted, flash frozen and stored at −80 °C until use.

For GELFrEE separation and LC–MS of integral membrane proteins, mitochondrial membrane proteins (400 μg as determined by bicinchoninic acid) were acetone precipitated, re-suspended in 100 μl loading buffer, reduced with 20 mM DTT and alkylated with 100 mM iodoacetamide. After a GELFrEE separation as reported previously[13], SDS was removed[36] and the fractions re-suspended in 30 μl of fresh 60% formic acid. Each 10-μl fraction was injected onto a PLRP analytical column (75 μm × 10 cm) and heated to 45 °C using a nanocapillary column heater (New Objective). A solvent system developed for integral membrane proteins was used in nanocapillary format here: A, 60% formic acid in water; B, 100% isopropanol. The gradient used was 0% B at 0 min, 25% B at 5 min, 60% B at 50 min, 95% B at 56 min. Ten GELFrEE samples were also analysed by the standard acetonitrile solvent and gradient system described above.

TMHMM version 2.0 (http://www.cbs.dtu.dk/services/TMHMM) was used for the transmembrane domain prediction of identified proteins. The SwissProt accession number for each identification was entered into the server to determine the number and location of all transmembrane domains. Differences between the database sequence and that identified by top-down mass spectrometry were accounted for by interrogating the identified sequence with TMHMM.

Protein identification and characterization. Mass spectrometry fragmentation data were acquired in three different modes, depending on protein mass range. Below 17 kDa, data-dependent collision-induced dissociation (CID) was used (FT/FT), whereas source-induced dissociation (SID) was used for masses between 17 and 25 kDa (FT/FT) and above 25 kDa (ion trap/FT). Based on preliminary analyses, SID of 15 V was optimal for ion trap scans for the dissociation of weakly bound non-covalent adducts, whereas the 75 V SID for fragmentation was standardized as described[15]. For data-dependent fragmentation (top two precursors, 15–25 m/z isolation window), dynamic exclusion was enabled with a repeat count of 2, an exclusion duration of 5,000 s, and a repeat duration of 240 s. Both CID and SID were collected using eight microscans and 85,000 resolving power (at $m/z = 400$) with a target value of 2 million charges. Data for H1299 cells obtained on the Orbitrap Elite were acquired using either higher-energy collisional dissociation, CID or electron transfer dissociation during top three data-dependent tandem mass spectrometry fragmentation using mass mode, four microscans, 108,000 resolving power and a target value of 1 million charges (cf. Fig. 4i, j, Supplementary Figs 12 and 13 and Supplementary Table 7).

Software and data analysis. Much of the software for intact mass determination (kDECON), generation of visual outputs (Proteome Display), and species differentiation (PTMcrawler) has been published recently[17]. Briefly, kDECON provides average mass information for charge state resolved distributions using the ion trap. PTMcrawler traverses a list of intact monoisotopic masses to find mass differences corresponding to PTMs such as methylations, acetylations and phosphorylations (Fig. 2b). Proteome Display was devised for visualization of nanocapillary LC–tandem mass spectrometry data, allowing graphical viewing of four-dimensional proteome runs and generation of plots for specific PTMs (Fig. 2).

The RAW files collected were first processed with an algorithm called CRAWLER to assign masses. Using a version of this program, scans collected by Fourier-transform mass spectrometry were de-isotoped using the Xtract or THRASH algorithm[38]. Scans collected in the ion trap were deconvoluted using the kDECON algorithm (minimum intensity cutoff 1,000; mass range 10–70 kDa). Xtract or THRASH processing generated monoisotopic neutral masses, whereas kDECON processing provided average neutral masses. Data-dependent fragmentation scans were summed within a retention time tolerance of 1 min and a precursor tolerance of 0.05 m/z whereas SID scans were summed in 0.3 min. In both cases, multiplexed fragmentation was considered. Fragmentation data were filtered by selecting the top three most intense neutral fragment masses within a 100-Da window below 2,000 Da, and the top five above 2,000 Da. One or more precursor masses and one or more fragment masses resulting from each summed unit were grouped as ProSightPC experiment, and written to a ProSight Upload Format file. If a precursor mass could not be determined, the experiment was written out to a separate file with a placeholder value, for separate analysis.

The ProSight Upload Format files output by CRAWLER were searched against a human proteome database using a custom implementation of ProSightPC 2.0 with iterative search logic[39–41] on a 168-node Rocks[42] cluster. Four types of analysis were run, depending on the type of data: FT/FT CID, FT/FT SID, ion trap/FT SID or SID data where the precursor could not be determined ('No-Hi-SID'). The iterative search trees were designed to take advantage of high mass accuracy, while retaining the option to run less specific searches if a result of sufficient quality could not be obtained by more specific searches. All searches used 10 p.p.m. tolerance for the fragment ions, all of which were obtained at high resolving power. For each search, the top ten hits were returned; if the top hit had an E value not more than 1×10^{-2}, the analysis moved on to the next experiment, otherwise the next search in the tree was run. All searches were in absolute mass mode. The FT–FT–CID tree consisted of searches at 200, 2,000 Da, and 'entire database' precursor tolerances; the FT–FT–SID at 2.3, 2,000, 20,000 and entire database; and ion trap–FT–SID trees used 2,000, 20,000 Da, and 'entire database' precursor tolerances; the No-Hi-SID tree just searched against the entire database. Searches were against two different human proteome databases built against UniProt release 2011_04, encompassing known alternative splices, modifications, peptide cleavage events, potential initial methionine cleavage and amino (N)-terminal acetylation. A complex database was created encompassing combinations of annotated alternative splice and peptide cleavage events to generate 54,190 base sequences. A maximum of 2^{13} protein forms for each base sequence make approximately 8,450,000 theoretical protein species. This database was used for all searches where the precursor tolerance was less than 2,000 Da. All other searches used a simplified database consisting of the same 54,190 base sequences, modified with N-terminal acetylation and initial methionine cleavage (where applicable) creating a total of approximately 160,000 forms.

Data were run against both forward and scrambled databases (see FDR estimation below) separately with identical search parameters. Upon completion, all search results were loaded into a ProSight data repository and a report was produced, returning the top hits for each experiment. The hit with the best q value was then chosen as the exemplar for each gene product cluster. If two member hits had the same q value, the member with the lowest absolute mass difference to the theoretical hit was chosen, and if this still produced a set with more than one member, the form with the 'most-terminal' PTMs (that is, closest to the N or carboxy terminus) was chosen. For a desired FDR cutoff (for example, 5%), a list of accession numbers and species is produced.

For estimation of FDR for top-down proteomics, the Poisson-based model as published in 2001 (ref. 43) had been previously modified with a Bonferroni correction that enables probability-based scoring (that is, use of the E value noted above) for searches done on a database created by shotgun annotation[16]. To validate the process used in high-throughput operations, an FDR analysis was performed to correct for multiple hypothesis testing using the method of Benjamini and Hochberg[44] as applied by Storey[45]. For each Poisson-based p value, a corresponding Bayesian posterior p value is calculated, termed the q value, which is a measure of the FDR for that particular identification event (also called an instantaneous FDR).

To calculate q values, a separate decoy database of scrambled sequences was created equal in size to the forward (real) database[46]. Searches were done separately on both the forward and decoy (scrambled) databases using all data in a set of four- or three-dimensional proteome runs. A histogram was created for the decoy database results using the log p_{id} value (where p_{id} is the Poisson probability of an incorrect protein identification) of these false identifications (Supplementary Fig. 14). These data were modelled against a gamma distribution and fit with a shape (k) of 10.26 ± 0.04 and rate (θ) of 2.25 ± 0.01. The distribution of scrambled hits is taken as an empirical estimate of the distribution of scores under the null hypothesis that the match was due to chance. Thus, the area under the scrambled score distribution to the right of the observed forward score is the probability of getting this good a forward score, or better, by chance (abbreviated hereafter p_F). From here, all data are rank-ordered by their corresponding p_F values and q values are calculated as in Storey[45]. The final results were generated using a q value cutoff of 0.05, thus achieving a protein level FDR of 5%. For comparison, bottom-up studies typically use a 1% FDR cutoff at the peptide level (the so-called PSM level, for peptide spectrum match), which typically rolls up into a 5–8% FDR at the protein level[47]. Extensive comparison with other FDR estimation techniques (including generating ROC curves for reversed decoy databases of concatenated reversed sequences) showed the q value approach to be 10–25% more stringent in terms of number of identified proteins and species.

31. Soubeyrand, S., Pope, L. & Hache, R. J. G. Topoisomerase II α-dependent induction of a persistent DNA damage response in response to transient etoposide exposure. *Mol. Oncol.* **4,** 38–51 (2010).
32. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16(INK4a). *Cell* **88,** 593–602 (1997).
33. Trinkle-Mulcahy, L. *et al.* Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol.* **183,** 223–239 (2008).
34. 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).
35. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of Bacteriopage T4. *Nature* **227,** 680–685 (1970).
36. Wessel, D. & Flugge, U. I. A Method for the Quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138,** 141–143 (1984).
37. Cox, B. & Emili, A. Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. *Nature Protocols* **1,** 1872–1878 (2006).
38. Horn, D. M., Zubarev, R. A. & McLafferty, F. W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **11,** 320–332 (2000).
39. Boyne, M. T. *et al.* Tandem mass spectrometry with ultrahigh mass accuracy clarifies peptide identification by database retrieval. *J. Proteome Res.* **8,** 374–379 (2009).
40. LeDuc, R. D. *et al.* ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* **32,** W340–W345 (2004).
41. Zamdborg, L. *et al.* ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **35,** W701–W706 (2007).
42. Papadopoulos, P. M., Katz, M. J. & Bruno, G. in *Proc. 3rd IEEE Int. Conf. Cluster Computing* 258 ⟨http://www.computer.org/portal/web/csdl/doi/10.1109/CLUSTR.2001.959986⟩ (2001).

43. Meng, F. Y. *et al.* Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nature Biotechnol.* **19,** 952–957 (2001).
44. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
45. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
46. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4,** 207–214 (2007).
47. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417 (2009).