Mol Cell Proteomics Papers in Press. Published on October 3, 2018 as Manuscript TIR118.000947

Quantifying homologous proteins and proteoforms

Dmitry Malioutov,[1] Tianchi Chen,[2] Edoardo Airoldi,[4] Jacob Jaffe,[3] Bogdan Budnik,[5] & Nikolai Slavov,[2,✉]

[1]T. J. Watson IBM Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
[2]Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
[3]Proteomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]Department of Statistics, Harvard University, Cambridge, MA 02138, USA
[5]MSPRL, FAS Division of Science, Harvard University, Cambridge, MA 02138, USA

✉ Correspondence should be addressed to: nslavov@alum.mit.edu

Many proteoforms – arising from alternative splicing, post-translational modifications (PTM), or paralogous genes – have distinct biological functions, such as histone PTM proteoforms. However, their quantification by existing bottom-up mass–spectrometry (MS) methods is undermined by peptide-specific biases. To avoid these biases, we developed and implemented a first-principles model (HI*quant*) for quantifying proteoform stoichiometries. We characterized when MS data allow inferring proteoform stoichiometries by HI*quant* and derived an algorithm for optimal inference. We applied this algorithm to infer proteoform stoichiometries in two experimental systems that supported rigorous bench-marking: alkylated proteoforms spiked-in at known ratios and endogenous histone 3 PTM proteoforms quantified relative to internal heavy standards. When compared to the benchmarks, the proteoform stoichiometries interfered by HI*quant* without using external standards had relative error of $5 - 15\%$ for simple proteoforms and $20 - 30\%$ for complex proteoforms. A HI*quant* server is implemented at: https://web.northeastern.edu/slavov/2014_HIquant/

1 Introduction

Alternative mRNAs splicing and post-translational modifications (PTM) produce multiple protein isoforms per gene, termed proteoforms by Smith et al. (1). Furthermore, protein isoforms can be produced by distinct but highly homologous open reading frames, i.e., paralogous genes. Despite having similar sequence, proteoforms and protein isoforms often have distinct, even opposite biological functions (2, 3). For examples: (i) some Bcl-x isoforms promote apoptosis while other Bcl-x isoforms inhibit apoptosis (4); (ii) the methylation of histone 3 can cause either transcriptional activation (lysine 4) or repression (lysine 9) depending on the modified lysine (5); (iii) pyruvate kinase isoforms have different metabolic regulation, activities, and roles in aerobic glycolysis (6–8); and (iv) minor amino acid differences between actin and tropomyosin isoforms alter the tropomyosin position on actin (9).

Understanding such systems demands accurate methods to distinguish and quantify proteoform abundances (3, 10). However, the high sequence homology among proteoforms makes their quantification by bottom-up LC-MS/MC very challenging since protease digestion (which is an essential part of high-throughput bottom-up proteomics) is likely to produce mostly peptides that are shared by multiples proteoforms (11–13). One approach to overcoming the challenge of shared peptides and distinguishing among proteoforms is to analyze the intact proteoforms without protease diges-

tion, which is know as top-down proteomics (10, 14–16). Combining advances in chromatographic separation of proteins with top-down proteomics has enabled quantifying functionally distinct isoforms of skeletal and cardiac muscle proteins (9, 17), combinatorial histone modifications (18, 19) and discovery analysis of thousands of human proteoforms (10, 20). Distinguishing proteoforms has been powered by efficient fragmentation methods, such as electron capture dissociation (ECD) (21) and electron transfer dissociation (ETD) (22). Combining these methods with infrared photoactivation (23) and with clever algorithms has made it possible to localize PTMs and identify the sequences and modified sites even for peptides with multiple covalent modifications (24, 25). Such methods, allow quantifying the ratios among modifications occurring on the same peptide.

While advances in protein separation and top-down methods are becoming increasingly powerful (10, 16, 20), bottom-up methods remain more widely used and still afford higher throughput (26, 27). When analyzed by bottom-up methods, many proteoforms-specific modifications occur on distant parts of the corresponding proteins and upon protease digestion will be found on different peptides. Furthermore, the majority of the peptides may be shared among proteoforms (11–13). Here we focus on quantifying the stoichiometry among such proteoforms from peptides quantified by bottom-up LC-MS/MS across multiple samples.

One validated approach to quantifying the stoichiometry among such proteoforms relies on external standards (28, 29). This method can afford high accuracy even for complex proteoforms (29). However, their wider use has been limited by expense and applied only to special cases that allow chemical modification of cell lysates, e.g., phosphorylation (30) and acetylation (31, 32). In the absence of external standards, the quantification of complex proteoform stoichiometries remains very challenging because the ratios between proteoform-specific peptides do not necessarily reflect the ratios between the corresponding proteoforms (33); precursor ion areas corresponding to the same phospho-site in the same sample can differ over 100-fold depending on the choice of protease (34). This discrepancy arises because a measured peptide level (precursor ion area) depends not only on the abundance of the corresponding protein(s) but also on extraneous factors including protein digestion, peptide ionization efficiency, the presence of other co-eluting peptides, and chromatographic aberrations (34–36). These extraneous factors break the equivalence between the abundance of a peptide and its precursor ion area and thus make protein quantification much more challenging than DNA quantification by sequencing. This problem is compounded when PTM peptides have been enriched, and thus their intensities scaled by unknown enrichment-dependent factors.

2 Experimental Procedures

We used two experimental approaches to derive bench-mark estimates of proteoform stoichiometries: (i) We mixed alkylated proteoforms of the dynamic universal proteomics standard (UPS2) into known ratios and (ii) We quantified stoichiometries among histone PTMs by parallel reaction monitoring (PRM) relative to peptide standards with known absolute abundances. Each of these methods is described in the subsections below, and the resulting bench-mark estimates of proteoform stoichiometries were used to evaluate HI*quant* inferences.

2.1 Quantifying alkylated proteoforms of UPS2

Our first test of HI*quant* on experimental data with bench-marked proteoforms was based on alkylated proteoforms of UPS2 that we mixed into predetermined ratios. To achieve such mixtures, we started by splitting a UPS standard into two equal parts, part A and part B. Then:

- Part A was reduced with TCEP, and Cys were alkylated with iodoacetamide.
- Part B was reduced with TCEP, and Cys were alkylated with vinylpyridine.

The alkylated proteoforms from both part A and part B were diluted to a total volume of $320\ \mu l$ and spiked into yeast samples according to the table below:

	TMT-126	TMT-127	TMT-128	TMT-129	TMT-130	TMT-131
Part A (iodoacetamide)	10	10	50	50	100	80
Part B (vinylpyridine)	50	100	40	50	10	50

This range of mixing ratios, spanning an order of magnitude, was chosen as it includes the range of relative changes commonly measured reliably by mass–spectrometry. The fractional site occupancy between the proteoforms varied within 10-fold in each direction, the range over which HI*quant* is likely to give the most accurate results. As discussed below, the ratios among proteoforms whose levels differ by many orders of magnitude may be better inferred from the precursor ion areas of their unique peptides.

After the spike in of UPS alkylated proteoforms, each sample was digested by trypsin overnight and labeled with the corresponding TMT tag according to manufacturers protocol. The labeled samples were mixed into a set and processed as described previously (8). Briefly, the set–sample

was injected from an auto–sampler into the trapping column ($75\ \mu m$ column ID, 5 cm length, and packed with $5\ \mu m$ beads with 20 nm pores; from Michrom Bioresources, Inc.) and washed for 15 min; the sample was eluted to analytic column (Waters columns with $75\ \mu m$ ID, 15 cm length, and packed with HSS T3 $1.8\ \mu m$ beads) with a gradient from 2 to 32 % of buffer B (0.1 % formic acid in ACN) over $180\ min$ gradient and fed into LTQ Orbitrap Elite (Thermo Fisher, San Jose, CA). The instrument was set to run in TOP 20 MS/MS mode method with dynamic exclusion. After MS1 scan in Orbitrap with 60K resolving power, each ion was submitted to an HCD MS/MS with 15K or 30K resolving power and to CID MS/MS scan subsequently. All quantification data were derived from HCD spectra.

In the particular case of UPS2, trypsin digestion resulted in enough peptides to constrain the HI*quant* model (Fig. 1) and to support accurate inference (Fig. 2b,c). More complex proteoforms may require more peptides than those quantified from a trypsin digestion. In such cases, to better constrain the inference, one might increase the number of peptides used by HI*quant* by compiling peptides quantified from several digestions, each of which using a different protease (34, 36). Since HI*quant* uses only the relative quantification and is insensitive to systematic biases, including proteoform-specific digestion efficiency, it can easily accommodate peptides quantified from digestions with different proteases or from different enrichment methods.

Analysis of mass-spec spectra

Mass/charge spectra were analyzed by MaxQuant (37) (version 1.4.1.2) or SEQUEST HT run via the Proteome Discoverer (64bit version 1.4.0.288, Thermo). All searches were run on a Windows server 2008 64 bit operating system with 64 CPU blades and 256 GB of RAM with the following general parameters. Parent ion mass tolerance was set to 20 ppm, mass tolerance for MS/MS ions was set to 0.02 Da for HCD and to 0.6 Da for CID spectra. For all searches, minimal peptide length was specified as 6 amino acids and maximal peptide length as 50 amino acids. The peptide charge state was limited to +7 for searches with MaxQuant. Searches were performed against the yeast uniprot database downloaded from www.yeastgenome.org in October 2014 containing 6,750 entries, the fasta sequences of the UPS standard, and common contaminants. Searches had trypsin specificity, allowing 2 missed cleavages. Asn and Gln deamidation and Met oxidation were included as variable modifications in the search parameters. There were no fixed modifications.

The search results from all search engines were filtered at 1 % false discovery rate (FDR) on both protein and on peptide levels using the Percolator (Version 2.05 Build Date May 6 2013). The results exported for further analysis included all peptide spectrum matches (PSM) that were assigned

to one or more proteins and passed the statistical significance filter. These results were outputted in the "evidence.txt" file for MaxQuant and in a peptide–level–results text file for Proteome Discoverer.

2.2 PRM of histone PTMs relative to heavy peptide standards

To evaluate HI*quant* inferences, we also used PTM stoichiometries estimated from a parallel reaction monitoring (PRM) assay quantifying each peptide relative to an internal heavy standard. Creech et al. have published a detailed description of the assay (29), and here we will only briefly summarize it. Drug-treated cells were collected by centrifugation. After cell lysis, histones were extracted with sulfuric acid and were precipitated with trichloroacetic acid. Samples comprised of 10 μg, a 5-fold reduction compared with Creech et al.(29), were propionylated, desalted, and digested overnight with trypsin. After a second round of propionylation, the samples were desalted using C18 Sep-Pak Cartridge (Waters). A mix of isotopically labeled synthetic peptides was spiked-in to each sample prior to MS analysis. Peptides were separated on a C18 column (EASY-nLC 1000, Thermo Scientific) and analyzed by MS in a PRM mode (Q Exactive-plus, Thermo Scientific) as described previously (29). Detailed SOPs for P100 and GCP assays, including synthetic peptide master mixture formulation, can be found at https://panoramaweb.org/labkey/wiki/LINCS/Overview%20Information/page.view?name=sops.

2.3 Inference model for HI*quant*

To infer proteoform stoichiometry, we use a simple model that is illustrated in Fig. 1a with proteoforms of histone H3 and in **Supplementary Fig. 1** with paralogous ribosomal proteins and phosphoproteoforms of pyruvate dehydrogenase. HI*quant* explicitly models peptide levels measured across conditions as a superposition of the levels of the proteins from which the peptides originate, Fig. 1a. In this model, shared peptides serve as indispensable internal standards; they couple the equations for different peptides and thus make possible estimating stoichiometries between homologous proteins and proteoforms. The simple example in Fig. 1a generalizes to any number of proteins / proteoforms (M) and any number of conditions greater than 1 ($N > 1$) as the system in Fig. 1b shows. HI*quant* solves this system and infers the protein levels (P) independently from the extraneous noise (Z; coming from protein-digestion, peptide-ionization differences, sample loss during enrichment, and even coisolation interference); Z is also inferred as part of the solution and discarded. A related superposition model has been used before with peptides quantified at one condition (13). However for a single condition, the model cannot quantify the proteins independently from the nuisance Z

since all problems described by system 1 in Fig. 1 are under-determined, i.e., have infinite number of solutions (Proof 1; Supplemental Information). Thus, for a single condition, the model cannot take advantage of the robust corresponding-ion pairs, i.e., ratios between ions with the same chemical composition. In contrast, HI*quant* infers ratios across proteins and their PTMs solely from the corresponding-ion ratios. This is possible because when $N > 1$, the system in Fig. 1b often has a unique solution up to a single scaling constant, even when all peptides are shared, e.g., the problem defined by the design matrix in **Supplementary Fig. 1**c. We characterize the conditions under which HI*quant* has a unique solution for the abundances of individual proteoforms and derive algorithms that use convex–optimization to find the optimal solution given the data; see Malioutov and Slavov (38) and Supplemental Information.

3 Validating inference of proteoform stoichiometry

Our model (Fig. 1b) aims to make proteoform quantification insensitive to many systematic biases. For example, incomplete cleavage of a peptide, e.g., only 5% of the peptide is released during enzyme digestion, is fully absorbed into the corresponding nuisance and does not affect inferred protein levels as long as the cleavage is 5% for all conditions/samples. Analogously, if coisolation interference compresses the fold-changes of a peptide, the systematic component of the compression is fully absorbed by the nuisances. Unlike systematic biases, random noise in the data is not absorbed by the nuisances; it can degrade the quality of the inference. In order to assess the reliability of the inferred proteoform abundances, HI*quant* carefully evaluates the inference and assigns confidence levels. The evaluation uses inference features, such as fraction of explained variance, eigenvalue spectrum spacing and noise sensitivity; see Supplemental Information.

We sought to experimentally evaluate HI*quant*'s ability to infer the proteoform stoichiometry in samples for which proteoform stoichiometries are accurately determined by other methods. The first method included creating and mixing alkylated proteoforms. The second method included quantifying histone H3 proteoforms relative to heavy peptide standards with known abundances.

3.1 Validation based on spiked-in alkylated proteoforms

We aimed to create proteoform mixtures with known stoichiometries so that they can be used to assess the accuracy of stoichiometries inferred by HI*quant*. To this end, the dynamic universal proteomics standard (UPS2) was split into two equal parts, A and B. In part A, cysteines were covalently modified with iodoacetamide, and in part B with vinylpyridine as described in the methods

and shown in Fig. 2a. We mixed part A and B in predefined ratios (n) and spiked each mixing ratio into an yeast sample. All samples were labeled with TMT, and the relative peptide levels quantified from the reporter ions at the MS2 level.

These alkylated UPS proteoforms have mostly shared peptides (peptides not containing cysteine) and only one or a few unique peptides per proteoform (peptides containing cysteine). HI*quant* modeled the relative levels of these peptides as shown in Fig. 1 and solved the model to infer the stoichiometries of the alkylated proteoforms ($\hat{n}$), which should correspond to the mixing ratios. Indeed, we find that the actual mixing ratios (n) for all quantified proteoforms correlate strongly to the inferred ratios ($\hat{n}$) as shown in Fig. 2b. To examine the accuracy of the inferred ratios $\hat{n}$ more closely, we sought to generate a distribution of errors between the inferred and the expected mixing ratios ($\hat{n}/n$) for many HI*quant* problems, more than the number of UPS2 proteoforms. To do so, we took advantage of the fact that the A / B ratios of the alkylated proteoforms for different UPS2 proteins should be the same since the proteins were mixed simultaneously and therefore in equal proportions. Thus, we created 1,500 HI*quant* problems by sampling with replacement shared peptides (containing cysteine) and unique peptides (not containing cysteine), and then inferred the $\hat{n}$ ratios between the iodoacetamide and vinylpyridine modified proteoforms, one ratios for each TMT channel and 6 per HI*quant* problem. The corresponding distribution of relative errors is shown in Fig. 2c along with the errors for ratios estimated from the abundances of the precursor ions. The median error is below 11 $\%$ for ratios inferred by HI*quant* and a substantially larger error for the ratios between precursor ion areas of unique peptides (Fig. 2c).

3.2 Validation based on histone 3 proteoforms quantified by PRM

Next, we sought to evaluate the ability of HI*quant* to infer stoichiometries of more complex PTM proteoforms, those of histone H3. We rigorously quantified endogenous proteoform stoichiometries by a previously developed assay based on external standards (MasterMix) with known concentrations (29). For the test, we used peptides quantified by parallel reaction monitoring (PRM) across 7 drug perturbations. Fractional site occupancies were estimated based on the external standards as described before by Creech et al (29). Independently, the same stoichiometries were inferred by HI*quant only* from the relative levels of the indigenous peptides, without using the MasterMix concentrations. The comparison of these estimates indicates good agreement (Fig. 3), supporting the ability of HI*quant* to infer fractional site occupancy even when the same site may be modified by different PTMs. The estimates from the external standards and from HI*quant* are very close

but also show some systematic deviations. Those deviations may arise due to incomplete protein digestion that is hard to control for with peptide standards, measurement noise corrupting the solution inferred by HI*quant* or proteforms not explicitly included in the model. The abundances of some proteoforms with quantified peptides is over 1000 fold lower than the abundance of the main proteoforms. They and their corresponding peptides were omitted from the HI*quant* inference since their quantification requires unrealistically high accuracy of relative quantification; see Supplemental Information and Discussion.

4 Discussion

The idea of using ratios between chemically identical ions is a cornerstone of quantitative proteomics (28, 39). It has been used for two decades in the context of relative quantification of proteins based on unique peptides (40) and even applied to the special case of inferring phosphorylation site occupancy (11, 33). Our work expands and generalizes this idea to all peptides, to stoichiometries of complex proteoforms, and to unlimited number of conditions. Crucially, HI*quant* allows accurate, efficient, and numerically stable inference resulting in reliability estimates.

HI*quant* requires and depends upon accurate relative quantification. This limitation is largely and increasingly mitigated by technological developments allowing accurate estimates of corresponding ion ratios. Such improvements include instrumentation advances (41, 42), interference free tandem mass tags (43), and enhanced peptide quantification from mass-spectra (27) However, these technological developments on their own do not allow accurate estimates of PTM site occupancy from bottom-up LC-MS/MS (34). HI*quant*'s dependence on the accuracy of relative quantification increases with increasing difference in the abundance of proteoforms. If the levels of two proteins differ by more than 3-6 orders of magnitude, this difference is likely better inferred from the precursor ion areas of the unique peptides. The associated noise (due to variability in protein digestion and ionization) is generally below 100 fold (36) and thus smaller than the signal. HI*quant*'s utility is particularly relevant when proteins and proteoforms have comparable abundances (within 10-100 fold difference) but distinct functions (44) and thus accurate quantification is essential for quantifying relatively small differences in abundance. Quantifying such proteoforms is an exciting frontier essential for understanding post-transcriptional regulation (45, 46) and defining cell-types from single cell proteomes (47).

The general form of HI*quant* described in Fig. 1c indicates that HI*quant* is not limited to proteoforms, even broadly defined. Rather, HI*quant* can be applied to any set of proteins sharing a

peptide. Here we emphasize the application to proteoforms because existing bottom-up methods are better suited for quantifying the stoichiometry between proteins with low homology that generate many unique peptides. For proteins with multiple unique peptides, some of the peptide-specific bias (from variation in protein-digestion and peptide-ionization efficiency) is likely to be averaged out and reduced. However, this bias is a more serious problem for proteoforms with only one or only a few unique peptides (34). For such proteoforms, HI*quant* can allow estimating stoichiometries accurately using only ratios between chemically identical ions.

Supplemental Information. Supplemental information includes Extended Experimental Procedures, Mathematical Proofs, and Supplemental Figures can be found in the Supplemental Information. The supplemental website for interactive data analysis can be found at:
https://web.northeastern.edu/slavov/2014_HIquant/
Python code implementing HI*quant* is available at: https://github.com/nslavov/HIquant

Acknowledgments. We thank S. Carr, A. Ivanov, S. MacNamara, Y. Katz, and MA. Blanco for help, critical discussions and feedback. N.S. started this work while a postdoc with Alexander van Oudenaarden and thanks him for generous support. Research was funded by a grants to N.S. from NIGMS of the NIH under Award Number DP2GM123497, a SPARC grant from the Broad Institute to N.S and S.C, and a NEU Tier 1 grant to N.S.

Data Availability: The data are available at both PRIDE (accession ID: PXD008557) and MassIVE (accession ID: MSV000081857).

References

1. Smith LM, Kelleher NL, et al. (2013) Proteoform: a single term describing protein complexity. Nature methods 10: 186–187.

2. Soria PS, McGary KL, Rokas A (2014) Functional divergence for every paralog. Molecular biology and evolution 31: 984–992.

3. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, et al. (2018) How many human proteoforms are there? Nature chemical biology 14: 206.

4. Schwerk C, Schulze-Osthoff K (2005) Regulation of apoptosis by alternative pre-mRNA splicing. Molecular cell 19: 1–13.

5. Berger SL (2007) The complex language of chromatin regulation during transcription. Nature 447: 407–412.

6. Tanaka T, Harano Y, Sue F, Morimura H (1967) Crystallization, characterization and metabolic regulation of two types of pyruvate kinase isolated from rat tissues. Journal of biochemistry 62: 71–91.

7. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, et al. (2008) The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature 452: 230–233.

8. Slavov N, Budnik B, Schwab D, Airoldi E, van Oudenaarden A (2014) Constant Growth Rate Can Be Supported by Decreasing Energy Flux and Increasing Aerobic Glycolysis. Cell Reports 7: 705 – 714.

9. Lehman W, Hatch V, Korman V, Rosol M, Thomas L, et al. (2000) Tropomyosin and actin isoforms modulate the localization of tropomyosin strands on actin filaments. Journal of molecular biology 302: 593–606.

10. Toby TK, Fornelli L, Kelleher NL (2016) Progress in top-down proteomics and the analysis of proteoforms. Annual Review of Analytical Chemistry 9: 499–519.

11. Dost B, Bandeira N, Li X, Shen Z, Briggs SP, et al. (2012) Accurate mass spectrometry based protein quantification via shared peptides. Journal of Computational Biology 19: 337–348.

12. Stastna M, Van Eyk JE (2012) Analysis of protein isoforms: can we do it better? Proteomics 12: 2937–2948.

13. Gerster S, Kwon T, Ludwig C, Matondo M, Vogel C, et al. (2014) Statistical approach to protein quantification. Molecular & Cellular Proteomics 13: 666–677.

14. Kelleher NL (2004) Top-down proteomics. Analytical Chemistry 76: 197A–203A.

15. Ge Y, Rybakova IN, Xu Q, Moss RL (2009) Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. Proceedings of the National Academy of Sciences 106: 12658–12663.

16. Siuti N, Kelleher NL (2007) Decoding protein modifications using top-down mass spectrometry. Nature methods 4: 817.

17. Sheng S, Chen D, Van Eyk JE (2006) Multidimensional liquid chromatography separation of intact proteins by chromatographic focusing and reversed phase of the human serum proteome optimization and protein database. Molecular & Cellular Proteomics 5: 26–34.

18. Young NL, DiMaggio PA, Plazas-Mayorca MD, Baliban RC, Floudas CA, et al. (2009) High throughput characterization of combinatorial histone codes. Molecular & Cellular Proteomics 8: 2266–2284.

19. Tian Z, Tolić N, Zhao R, Moore RJ, Hengel SM, et al. (2012) Enhanced top-down characterization of histone post-translational modifications. Genome biology 13: R86.

20. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, et al. (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature 480: 254.

21. Zubarev RA, Kelleher NL, McLafferty FW (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. Journal of the American Chemical Society 120: 3265–3266.

22. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 101: 9528–9533.

23. Riley NM, Sikora JW, Seckler HS, Greer JB, Fellers RT, et al. (2018) The value of activated ion electron transfer dissociation for high-throughput top-down characterization of intact proteins. Analytical Chemistry .

24. DiMaggio PA, Young NL, Baliban RC, Garcia BA, Floudas CA (2009) 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. Molecular & Cellular Proteomics 8: 2527–2543.

25. Guan S, Burlingame AL (2010) Data processing algorithms for analysis of high resolution MSMS spectra of peptides with complex patterns of posttranslational modifications. Molecular & Cellular Proteomics 9: 804–810.

26. Gillet LC, Leitner A, Aebersold R (2016) Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. Annual review of analytical chemistry 9: 449–472.

27. Sinitcyn P, Rudolph JD, Cox J (2018) Computational methods for understanding mass spectrometry–based shotgun proteomics data. Annu Rev Biomed Data Sci 1: 207–34.

28. Ong SE, Mann M (2005) Mass spectrometry–based proteomics turns quantitative. Nature chemical biology 1: 252.

29. Creech AL, Taylor JE, Maier VK, Wu X, Feeney CM, et al. (2015) Building the connectivity map of epigenetics: Chromatin profiling by quantitative targeted mass spectrometry. Methods 72: 57–64.

30. Wu R, Haas W, Dephoure N, Huttlin EL, Zhai B, et al. (2011) A large-scale method to measure absolute protein phosphorylation stoichiometries. Nature methods 8: 677–683.

31. Weinert BT, Iesmantavicius V, Moustafa T, Schölz C, Wagner SA, et al. (2014) Acetylation dynamics and stoichiometry in saccharomyces cerevisiae. Molecular Systems Biology 10.

32. Baeza J, Dowell JA, Smallegan MJ, Fan J, Amador-Noguez D, et al. (2014) Stoichiometry of site-specific lysine acetylation in an entire proteome. Journal of Biological Chemistry 289: 21326–21338.

33. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, et al. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Science signaling 3: ra3–ra3.

34. Giansanti P, Aye TT, van den Toorn H, Peng M, van Breukelen B, et al. (2015) An augmented multiple-protease-based human phosphopeptide atlas. Cell reports 11: 1834–1843.

35. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2006) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature biotechnology 25: 117–124.

36. Peng M, Taouatas N, Cappadona S, van Breukelen B, Mohammed S, et al. (2012) Protease bias in absolute protein quantitation. Nature methods 9: 524–525.

37. Cox J, Mann M (2008) Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature biotechnology 26: 1367–1372.

38. Malioutov D, Slavov N (2014) Convex Total Least Squares. Journal of Machine Learning Research 32: 109 – 117.

39. Blagoev B, Ong SE, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nature biotechnology 22: 1139–1145.

40. Altelaar A, Frese CK, Preisinger C, Hennrich ML, Schram AW, et al. (2013) Benchmarking stable isotope labeling based quantitative proteomics. Journal of proteomics 88: 14–26.

41. Zubarev RA, Makarov A (2013) Orbitrap mass spectrometry. Analytical Chemistry 85: 5288–5296.

42. Nagornov KO, Kozhinov AN, Tsybin YO (2017) Fourier transform ion cyclotron resonance mass spectrometry at the cyclotron frequency. Journal of The American Society for Mass Spectrometry 28: 768–780.

43. Winter SV, Meier F, Wichmann C, Cox J, Mann M, et al. (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. Nature methods : doi: 10.1038/s41592–018–0037–8.

44. Slavov N, Semrau S, Airoldi E, Budnik B, van Oudenaarden A (2015) Differential stoichiometry among core ribosomal proteins. Cell Reports 13: 865 – 873.

45. van den Berg PR, Budnik B, Slavov N, Semrau S (2017) Dynamic post-transcriptional regulation during embryonic stem cell differentiation. bioRxiv 1: doi: 10.1101/123497.

46. Franks A, Airoldi E, Slavov N (2017) Post-transcriptional regulation across human tissues. PLoS computational biology 13: e1005535.

47. Budnik B, Levy E, Harmange G, Slavov N (2017) Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. bioRxiv 1: DOI: 10.1101/102681.

Figure 1

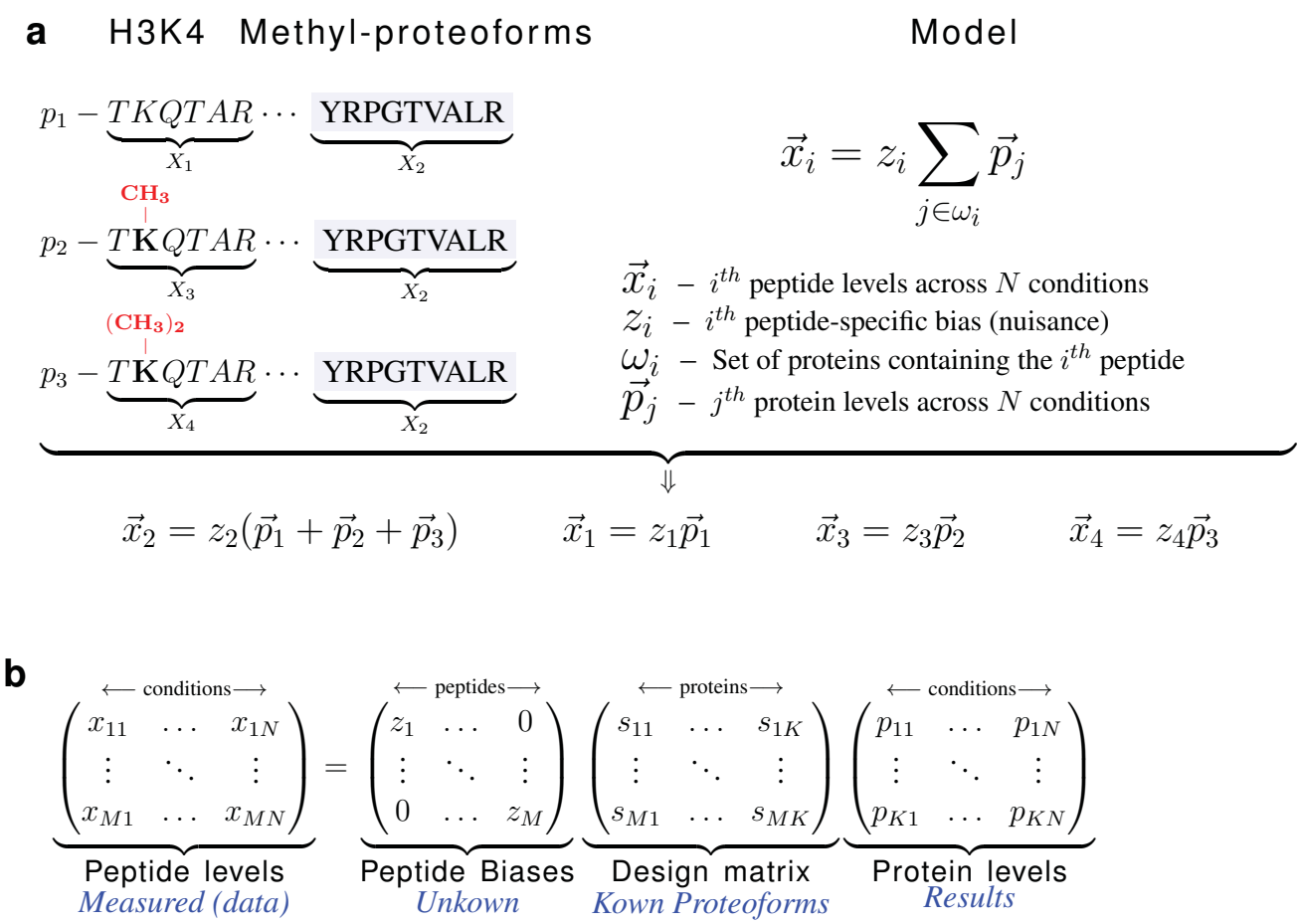

Figure 1 | Model for inferring stoichiometries among proteoforms and paralogous proteins independently from peptide-specific biases. (**a**) One shared (X_2) and three unique (X_1, X_3 and X_4) peptides of H3 proteoforms illustrate a very simple case of HI*quant*. HI*quant* models the peptide levels measured across conditions ($\vec{x}$) as a supposition of the protein levels ($\vec{p}$), scaled by unknown peptide–specific biases/nuisances (z). These coupled equations can be written in a matrix form whose solution infers the methylation stoichiometry independently from the nuisances (z). (**b**) The general form of the model for K proteoforms (or homologous proteins) with M peptides quantified across N conditions can be formulated and solved. In many, albeit not all, cases an optimal and unique solution can be found, even in the absence of unique peptides; see **Supplementary Fig. 1** and Supplemental Information.

Figure 2

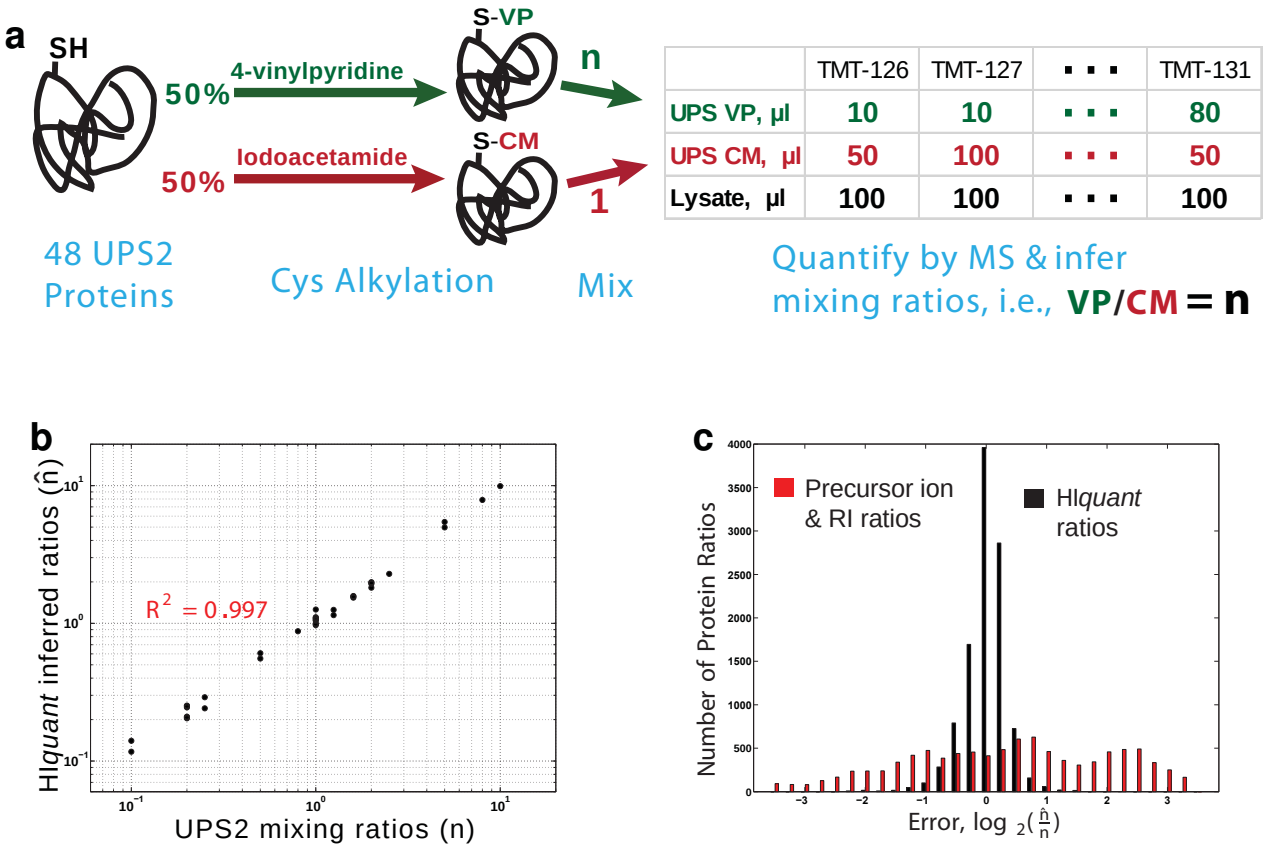

Figure 2 | HI*quant* accurately quantifies ratios across alkylated proteoforms of a spiked-in standard. (**a**) Schematic diagram of a validation experiment. We prepared a gold standard of proteoforms from the dynamic universal proteomics standard (UPS2) whose cysteines were covalently modified either with iodoacetamide or with vinylpyridine. Upon digestion, these modified UPS proteins generate many shared peptides (peptides not containing cysteine) and a few unique peptides (peptides containing cysteine). The modified UPS2 proteins were mixed with one another at known ratios (n), mixed with yeast lysate, digested and quantified by MS. The proteoform ratios that HI*quant* inferred from the MS data ($\hat{n}$) were compared to the mixing ratios. (**b**) The ratios *across* the alkylated isoforms of UPS2 inferred by HI*quant* ($\hat{n}$, y-axis) accurately reflect the mixing ratios (n, x-axis). (**c**) The mixing and inferred ratios in panel b span 2-orders of magnitude, which is much larger than the dynamic range of relative error. To zoom in on the relative errors, we plotted a distribution of $\log_2(\frac{\hat{n}}{n})$ for $1,500$ HI*quant* problems generated by sampling with replacement peptides from all UPS2 proteins. For HI*quant*, this distribution indicates small error, with median error below 11%. However the ratios estimated just from the precursor intensities of the unique peptides for each proteoform show significantly higher relative error, mostly likely because of peptide-specific variability in digestion and ionization.

Figure 3

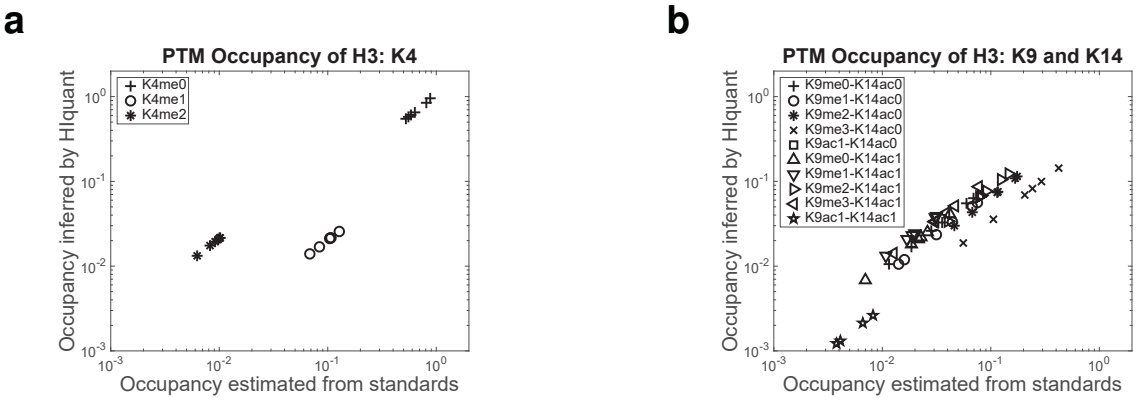

Figure 3 | HI*quant* accurately infers stoichiometries and confidence intervals across PTM site occupancies of histone 3. (**a**) Histone 3 peptides were quantified by SRM across 7 perturbations, and the fractional site occupancies for K4 methylation estimated by two methods: Estimates inferred by HI*quant* without using external standards are plotted against the corresponding estimates based on MasterMix external standards with known concentrations (29). Each marker shape corresponds to the PTM site(s) shown in the legend; methylation is denoted with "me" and acetylation with "ac" followed by the number of methyl/acetyl groups. (**b**) The validation method from (a) was extended to another set of more complex fractional site occupancies on K9 methylation and K14 acetylation.