Quantifying homologous proteins and proteoforms

Dmitry Malioutov,[1] Tianchi Chen,[2] Jacob Jaffe,[3] Edoardo Airoldi,[4] Steve Carr,[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]FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA

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

Many proteoforms – arising from alternative splicing, post-translational modifications (PTMs), 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*, derived an algorithm for optimal inference, and demonstrated experimentally high accuracy in quantifying fractional PTM occupancy without using external standards, even in the challenging case of the histone modification code. A HI*quant* server is implemented at: https://web.northeastern.edu/slavov/2014_HIquant/

1 Introduction

Alternative mRNAs splicing and post-translational modifications (PTMs) produce multiple protein isoforms per gene, termed proteoforms by Smith et al. (2013). 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 (Soria et al., 2014). For examples: (i) some Bcl-x isoforms promote apoptosis while other Bcl-x isoforms inhibit apoptosis (Schwerk and Schulze-Osthoff, 2005); (ii) the methylation of histone 3 can cause either transcriptional activation (lysine 4) or repression (lysine 9) depending on the modified lysine (Berger, 2007); and (iii) pyruvate kinase isoforms have different metabolic regulation, activities, and roles in aerobic glycolysis (Tanaka et al., 1967; Christofk et al., 2008; Slavov et al., 2014).

Understanding such systems demands quantifying proteoform abundances. This demand has motivated the development of external standards that can afford high accuracy even for complex proteoforms (Creech et al., 2015). 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 (Wu et al., 2011) and acetylation (Weinert et al., 2014; Baeza et al., 2014). 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 (Olsen et al., 2010); precursor ion areas corresponding to

the same phospho-site in the same sample can differ over 100-fold depending on the choice of protease (Giansanti et al., 2015). This is 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 (Lu et al., 2006; Peng et al., 2012; Giansanti et al., 2015). 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 Results

2.1 Model

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 phospho-proteoforms 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 (Gerster et al., 2014). However for a single condition, the model cannot quantify the proteins independently from the nuisance Z since all problems described by system 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 and derive algorithms that use convex–optimization to find the optimal solution given the data; see Malioutov and Slavov (2014) and Supplemental Information.

2.2 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. Thus, HI*quant* must carefully evaluate the reliability of interfered proteoform 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 stoichiometry is accurately determined by other methods. The first method included creating and mixing proteoforms. The second method included quantifying histone H3 proteoforms relative to heavy peptide standards with predetermined abundances.

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 digested, and the peptides split into two equal parts, A and B. In part A, cysteines were covalently modified with iodoacetamide, and in part B with vinylpyridine, 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 a few unique peptides (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. A comparison between the actual mixing ratios (n) with the inferred ratios ($\hat{n}$) demonstrates a median error below 10 %

for ratios inferred by HI*quant* and a substantially larger error for the ratios between precursor ion areas of unique peptides (Fig. 2b,c); see Supplemental Information.

Next, we sought to evaluate the ability of HI*quant* to infer stoichiometries of more complex PTM proteoforms, those of histone H3. This system allows rigorous quantification of endogenous proteoform stoichiometries by a previously developed external standards (MasterMix) with known concentrations (Creech et al., 2015). For the test, we used peptides quantified by selective reaction monitoring (SRM) across 7 perturbations. Fractional site occupancies were either estimated based on the external standards or inferred by HI*quant only* from the relative levels of the indigenous peptides, without using the MasterMix concentrations. The good agreement between these estimates (Fig. 3) validates 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.

3 Discussion

The idea of using ratios between chemically identical ions is a cornerstone of quantitative proteomics (Blagoev et al., 2004). It has been used for decades in the context of relative quantification of proteins based on unique peptides (Altelaar et al., 2013) and even applied to the special case of inferring phosphorylation cite occupancy (Olsen et al., 2010). 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. However, these technological developments on their own do not allow accurate estimates of PTM site occupancy (Giansanti et al., 2015). 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 (Peng et al., 2012) 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 (Slavov et al., 2015) and thus accurate quantification is essential for quantifying relatively small differences in abundance. Quantifying proteoforms is an exciting frontier essential for understanding post-transcriptional regulation (Floor and Doudna, 2016; Franks et al., 2017) and defining cell-types from single cell proteomes (Budnik et al., 2017).

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 (Giansanti et al., 2015). 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/

Acknowledgments. We thank M. Jovanovic, S. MacNamara, Y. Katz, and MA. Blanco for 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.

References

Altelaar, A., Frese, C. K., Preisinger, C., Hennrich, M. L., Schram, A. W., Timmers, H., Heck, A. J. and Mohammed, S. (2013). Benchmarking stable isotope labeling based quantitative proteomics. Journal of proteomics *88*, 14–26.

Baeza, J., Dowell, J. A., Smallegan, M. J., Fan, J., Amador-Noguez, D., Khan, Z. and Denu, J. M. (2014). Stoichiometry of site-specific lysine acetylation in an entire proteome. Journal of Biological Chemistry *289*, 21326–21338.

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. Nature *447*, 407–412.

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

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

Christofk, H. R., Vander Heiden, M. G., Harris, M. H., Ramanathan, A., Gerszten, R. E., Wei, R., Fleming, M. D., Schreiber, S. L. and Cantley, L. C. (2008). The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature *452*, 230–233.

Creech, A. L., Taylor, J. E., Maier, V. K., Wu, X., Feeney, C. M., Udeshi, N. D., Peach, S. E., Boehm, J. S., Lee, J. T., Carr, S. A. et al. (2015). Building the Connectivity Map of epigenetics: Chromatin profiling by quantitative targeted mass spectrometry. Methods *72*, 57–64.

Floor, S. N. and Doudna, J. A. (2016). Tunable protein synthesis by transcript isoforms in human cells. Elife *5*, e10921.

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

Gerster, S., Kwon, T., Ludwig, C., Matondo, M., Vogel, C., Marcotte, E. M., Aebersold, R. and Bühlmann, P. (2014). Statistical approach to protein quantification. Molecular & Cellular Proteomics *13*, 666–677.

Giansanti, P., Aye, T. T., van den Toorn, H., Peng, M., van Breukelen, B. and Heck, A. J. (2015). An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. Cell reports *11*, 1834–1843.

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

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

Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A. et al. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Science signaling *3*, ra3–ra3.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A. and Heck, A. J. (2012). Protease bias in absolute protein quantitation. Nature methods *9*, 524–525.

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

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

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

Smith, L. M., Kelleher, N. L. et al. (2013). Proteoform: a single term describing protein complexity. Nature methods *10*, 186–187.

Soria, P. S., McGary, K. L. and Rokas, A. (2014). Functional divergence for every paralog. Molecular biology and evolution *31*, 984–992.

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

Weinert, B. T., Iesmantavicius, V., Moustafa, T., Schölz, C., Wagner, S. A., Magnes, C., Zechner, R. and Choudhary, C. (2014). Acetylation dynamics and stoichiometry in *Saccharomyces cerevisiae*. Molecular systems biology *10*, 716.

Wu, R., Haas, W., Dephoure, N., Huttlin, E. L., Zhai, B., Sowa, M. E. and Gygi, S. P. (2011). A large-scale method to measure absolute protein phosphorylation stoichiometries. Nature methods *8*, 677–683.

Figure 1

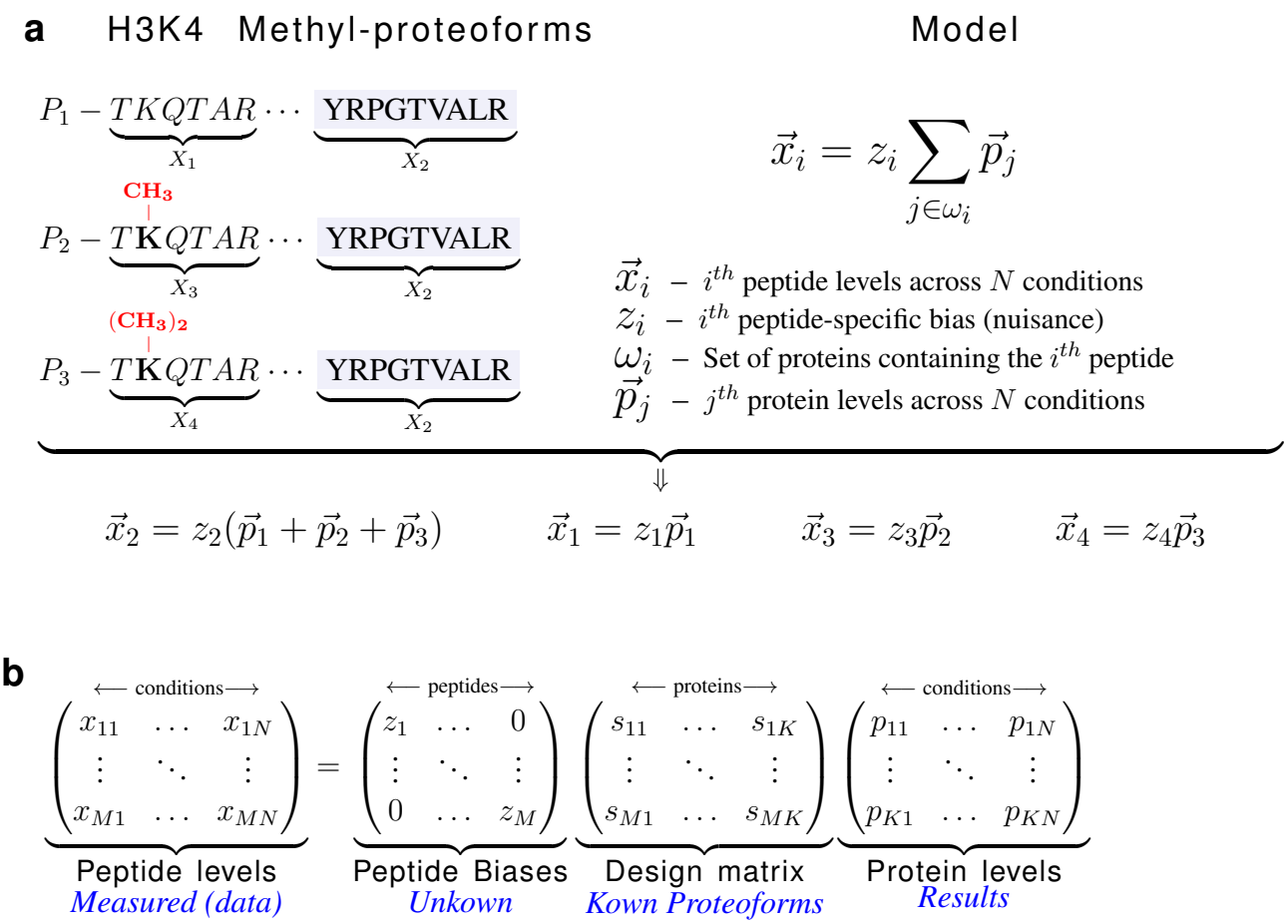

Figure 1 | Model for inferring stoichiometries among proteoforms and paralogous proteins independently from peptide-specific biases. (**a**) One shared (X_2) and three unique (X_1, X_3 and X_4) peptides of H3 proteoforms illustrate a very simple case of HI*quant*. HI*quant* models the peptide levels 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

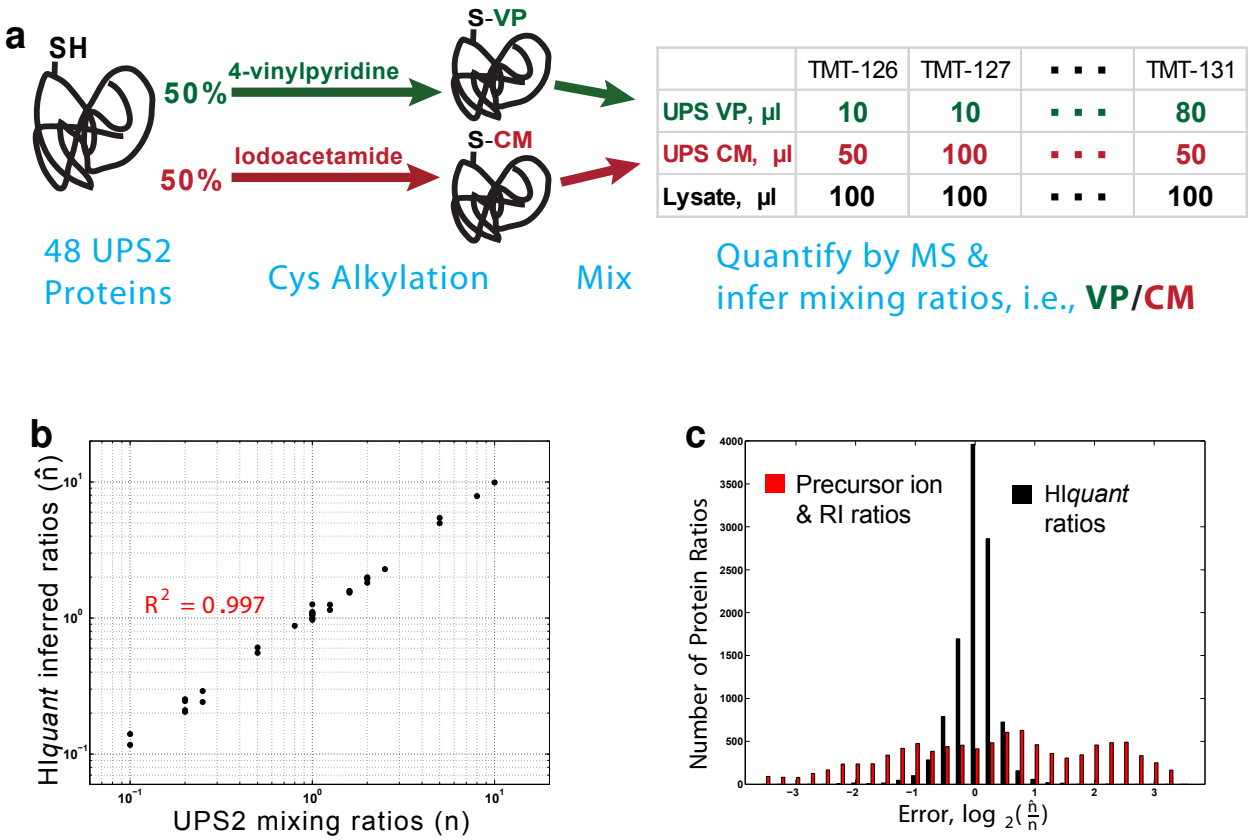

Figure 2 | HI*quant* accurately quantifies ratios across alkylated proteoforms of a spiked-in standard. (**a**) Schematic diagram of a validation experiment. We prepared a gold standard of proteoforms from the dynamic universal proteomics standard (UPS2) whose cysteines were covalently modified either 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**) Comparison of the error in proteoform ratios inferred by HI*quant* and ratios inferred from the precursor ion areas and the reporter ion (RI) ratios.

Figure 3

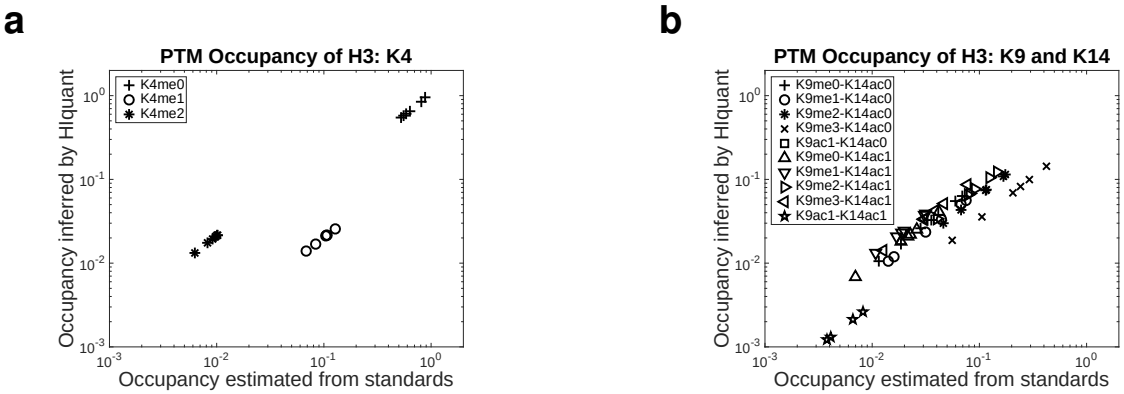

Figure 3 | HI*quant* accurately infers stoichiometries and confidence intervals across PTM site occupancies of histone 3. (**a**) Histone 3 peptides were quantified by SRM across 7 perturbations, and the fractional site occupancies for K4 methylation estimated by two methods: Estimates inferred by HI*quant* without using external standards are plotted against the corresponding estimates based on MasterMix external standards with known concentrations (Creech et al., 2015). 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.

Supplemental Figure 1

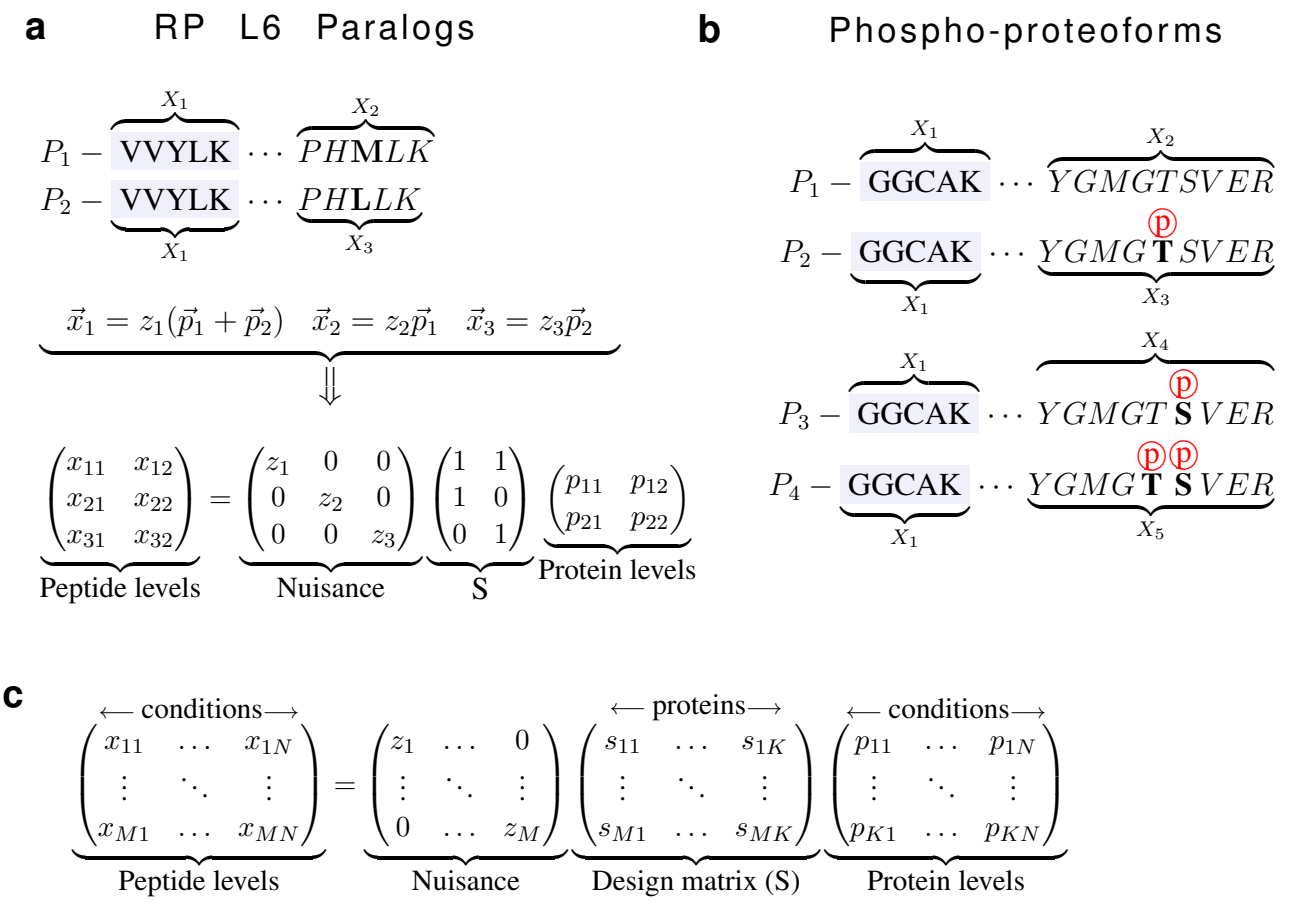

Figure S1 | Model for inferring stoichiometries among proteoforms and paralogous proteins independently from peptide-specific biases. (**a**) One shared (X_1) and two unique (X_2 and X_3) peptides from the two paralogs of ribosomal proteins L6 illustrate the simplest case of HI*quant*. HI*quant* models the peptide levels measured across two conditions ($\vec{x}$) as a supposition of the protein levels ($\vec{p}$), scaled by unknown peptide–specific nuisances (z). These coupled equations can be written in a matrix form whose solution infers the P_1/P_2 stoichiometry independently from the nuisances (z). (**b**) The shared and unique peptides of proteoforms (as illustrated by PDHA1 phospho-proteoforms) can be modeled as in panel (a); (**c**) The matrix system from (a) generalizes to K proteoforms (and homologous proteins) with M peptides quantified across N conditions. In many, albeit not all, cases an optimal and unique solution can be found, even in the absence of unique peptides. See Supplemental Information for details.