Journal of proteome research

Article

pubs.acs.org/jpr

Analysis of Intrinsic Peptide Detectability via Integrated Label-Free and SRM-Based Absolute Quantitative Proteomics

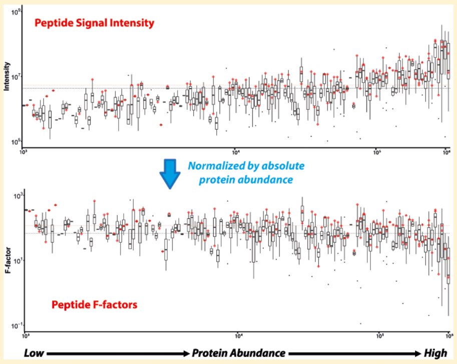
Andrew F. Jarnuczak,[†] Dave C. H. Lee,[‡] Craig Lawless,[†] Stephen W. Holman,[§] Claire E. Eyers,[*,§] and Simon J. Hubbard[*,†]

[†] Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom

[‡] Faculty of Biology, Medicine and Health, University of Manchester, Second Floor, Wolfson Molecular Imaging Centre, 27 Palatine Road, Withington, Manchester, M20 3JL, United Kingdom

[§] Centre for Proteome Research, University of Liverpool, Department of Biochemistry, Institute of Integrative Biology, Crown Street, Liverpool, L69 7ZB, United Kingdom

S *Supporting Information*



ABSTRACT: Quantitative mass spectrometry-based proteomics of complex biological samples remains challenging in part due to the variability and charge competition arising during electrospray ionization (ESI) of peptides and the subsequent transfer and detection of ions. These issues preclude direct quantification from signal intensity alone in the absence of a standard. A deeper understanding of the governing principles of peptide ionization and exploitation of the inherent ionization and detection parameters of individual peptides is thus of great value. Here, using the yeast proteome as a model system, we establish the concept of peptide F-factor as a measure of detectability, closely related to ionization efficiency. F-factor is calculated by normalizing peptide precursor ion intensity by absolute abundance of the parent protein. We investigated F-factor characteristics in different shotgun proteomics experiments, including across multiple ESI-based LC−MS platforms. We show that F-factors mirror previously observed physicochemical predictors as peptide detectability but demonstrate a nonlinear relationship between hydrophobicity and peptide detectability. Similarly, we use F-factors to show how peptide ion coelution adversely affects detectability and ionization. We suggest that F-factors have great utility for understanding peptide detectability and gas-phase ion chemistry in complex peptide mixtures, selection of surrogate peptides in targeted MS studies, and for calibration of peptide ion signal in label-free workflows. Data are available via ProteomeXchange with identifier PXD003472.

KEYWORDS: *quantitative proteomics, peptide ionization, peptide detectability, hydrophobicity, coelution*

INTRODUCTION

Building on advances in modern mass spectrometry, proteomics as a science has arguably come of age.[1] As a valuable tool in the modern postgenomic era, it seeks to characterize the complete set of protein molecules encoded and expressed by a genome in a cell or tissue under defined conditions. In other words, it should be truly genome-wide, and this is reportedly now readily achievable (e.g., refs 2−5). Equally important, for protein characterization to be truly informative, such an analysis should be quantitative. Indeed, there are now many instrument platforms that support high-coverage quantitative proteomics capable of generating estimates of both relative and absolute abundance of individual proteins in a complex mixture.[3,6,7] Ideally, such quantitative information should be absolute, i.e., yield a measure of concentration or copies per cell, as this empirically supports more thorough and biologically relevant analyses including determining the stoichiometry of protein complexes, biomarker assays, and measurement of protein concentrations for kinetic modeling.[8] However, estimating absolute protein abundance from peptide precursor ion signal requires calibration using known quantities of an internal standard.[7,9,10] Such calibration in quantitative proteomics is essential because the relative signal intensity arising from different peptide ions following electrospray ionization is not directly proportional to the amount (moles) of starting material; proteolytic peptides derived from the same parent protein, even when present in stoichiometric amounts, will not generate the same number of ions and thus their attendant MS signal will vary. Because the peptide precursor ion intensity (MS1) is frequently used for quantification, it is perhaps somewhat surprising that more attention is not generally given to evaluating the detectability of the proteolytic peptides used for quantification and the physicochemical factors governing these processes.

Received: January 19, 2016

DOI: 10.1021/acs.jproteome.6b00048

There are numerous confounding factors contributing to the detection of peptide ions in a mass spectrometer.[11] First, as noted, the intrinsic ionization properties of peptides vary dependent on the constituent amino acids and their order. Equally, although not essential for quantification per se, sufficient fragmentation should occur to allow search tools to accurately assign a candidate identification from within a complex mixture because in practice the MS1 peptide precursor ion signal needs to be assigned to the parent protein for quantification. Another confounding issue is the difference and nonuniformity in the chromatographic system delivering peptides to the mass spectrometer in most experiments, where different gradients and columns can cause differential peptide coelution and potentially lead to variation in competition for ionization. Similarly, missed or nonspecific cleavages by the protease employed to generate the peptides in the first instance can induce further variance. The field has thus invested much time and effort to the optimization of digestion protocols to induce complete proteolytic cleavage.[12−14] Biological and chemical modification can also split signal across multiple peptide ion species, leading to challenges when estimating protein abundance from the MS ion signal.

Hence, there is great value in understanding the issues surrounding intrinsic peptide “flyability” (a generic term used to cover the relative efficiencies of ionization, transfer, and detection) for several reasons; a greater understanding of the fundamental properties of gas-phase peptide ion chemistry, better selection of surrogate peptides for protein quantification experiments (because confounding, nonquantotypic peptides will not be helpful), and for more accurate estimation of absolute protein abundance from label-free data sets where calibration can be achieved from stable-isotope standards or intrinsic scaling factors.[7,9,15,16] Although the mechanistic principles of peptide fragmentation by collisional-based dissociation have been studied extensively and are reasonably well understood,[17−21] the basic principles of peptide ionization, particularly by electrospray and in complex mixtures, are still not fully deciphered despite numerous previous studies.[22−27] A deeper understanding of peptide detectability has particular relevance for quantitative proteomic strategies, which seek to predict quantotypic peptides and their likely signal strength for the selection of surrogate peptides or to calibrate precursor ion signals.[28−32]

To address this issue, we have exploited a data set of high-quality absolute quantifications of 349 yeast proteins derived from targeted MS experiments; selected reaction monitoring (SRM)[33] coupled to stable isotope-labeled standards in the form of QconCATs[34,35] was used to define accurate and reliable copies per cell abundance values.[36] This data was in turn used to normalize the peptide ion intensities captured in several parallel label-free shotgun studies carried out on the same yeast protein lysate.

By accounting for the major source of variance in the peptide intensity signal, namely, differential abundance of the peptides in the starting analyte mixture, we aim to get a closer estimate of the intrinsic peptide detectability, which we define here as individual peptide “flyability” factors (F-factors). F-factors are hence a more useful metric when examining peptide properties related to ionization efficiency because they effectively normalize for protein abundance as a confounding feature contributing to differences in relative ion signal intensity and the knock-on ability to identify the peptide.[11,37] Here, we calculate F-factors for thousands of yeast peptides in data-dependent label-free proteomics experiments and examine their statistical properties. We show that F-factors are, to a large degree, well conserved between experimental conditions but less so between different instrument types. Finally, we discuss the general implications for protein quantification via label-free peptide ion signals.

■ MATERIALS AND METHODS

Sample Preparation and Proteolysis

S. cerevisiae (EUROSCARF accession number Y11335 BY4742; Mat ALPHA; his3Δ1; leu2Δ0; lys2Δ0; ura3Δ0; YJL088w::kanMX4) cultures were prepared as described previously.[35,36,38] Briefly, four biological replicates of yeast cultures grown in carbon-limited medium, in chemostat mode, were collected, and the total number of harvested cells was determined using an automated cell counter (Cellometer AUTOM10).

For protein extraction, cell pellets were resuspended in 50 mM ammonium bicarbonate containing protease inhibitors cocktail (Roche Diagnostics Ltd., West Sussex, UK) and subjected to 15 cycles of bead-beating. Total protein extracts were collected, and protein concentration was determined using a Bradford assay.

Proteins corresponding to 25 million cells (∼100 μg protein) were digested using the procedure described in ref 38. Briefly, yeast lysate, universal proteomics standard (UPS1; Sigma-Aldrich), or proteomics dynamic range standard (UPS2; Sigma-Aldrich) in 25 mM ammonium bicarbonate was denatured using 1% (w/v) RapiGest (Waters, Manchester, UK), reduced by the addition of 60 mM dithiothreitol, and alkylated using 180 mM iodoacetamide (final concentrations were 0.05%, 3 mM, and 9 mM, respectively). Digestion was performed by the addition of trypsin at a 1:50 enzyme to protein ratio followed by another aliquot after 4.5 h and overnight incubation at 37 °C. The digestion was stopped, and RapiGest was removed by acidification with trifluoroacetic acid to a final concentration of 1% (v:v). Additionally, a 7.5 μL aliquot of acetonitrile:water (2:1) was added to aid peptide solubilization. RapiGest and any remaining cell debris were then removed by centrifugation (15000*g* for 20 min). Digested UPS standards were diluted to a final protein concentration of 25 fmol/μL (UPS1 containing equimolar amounts of 48 human proteins) or ranging from 250 fmol/μL to 2.5 amol/μL (UPS2 containing 48 human proteins spanning a concentration range of 5 orders of magnitude). Each standard was spiked into the *S. cerevisiae* lysate to an appropriate concentration to achieve 25 fmol (or 250 fmol to 2.5 amol range) standard in 500 ng yeast on column. Peptides were stored at −20 °C prior to MS analysis.

MassPREP *E. coli* digestion standard (Product Number 186003196; Waters, Manchester, UK), which is a tryptic digest of a purified *E. coli* cytosolic protein fraction, was solubilized in water:acetonitrile (97:3 with 0.1% (v/v) trifluoroacetic acid), aliquoted, and stored at −20 °C. For the MS analysis, 500 ng of yeast and 500 ng of MassPREP *E. coli* standard were mixed and injected onto the column.

Mass Spectrometry Data Acquisition and Processing

Protein digests were separated by reversed-phase liquid chromatography and analyzed on a LTQ-Orbitrap Velos equipped with a Nanospray Flex Ion Source or a Q Exactive HF mass spectrometer equipped with an EASY-Spray Source. For the LTQ-Orbitrap Velos, a nanoAcquity UPLC system (Waters, Manchester, UK) with a 75 μm × 25 cm, 1.8 μm particle size, C18 nanoAcquity analytical column was used. For

DOI: 10.1021/acs.jproteome.6b00048

Table 1. Primary Experimental Datasets Used in This Study

data set	instrument	gradient	theoretical peptides	peptide identifications (FDR < 0.01)	number of F-factors calculated	matching Q-peptides observed[a]
Yeast_HF_60	QEx-HF	60 min	~200k	16,078	3249	402
Yeast_HF_120	QEx-HF	120 min	~200k	23,383	4142	485
Yeast_Velos_240	LTQ-Velos	240 min	~200k	12,581	2773	347
Yeast_Velos_50	LTQ-Velos	30 min (50 min run)	~200k	3,810	983	137

[a] Number of peptides identified in a given run that were also used for quantification in the QConCAT study.[36].

the Q Exactive HF, Dionex UltiMate 3000 ultrahigh pressure LC system (Thermo Fisher Scientific, Hemel Hempstead, UK) with a 75 μm × 50 cm, 2 μm particle size, EASY-Spray analytical column was used. Peptides were loaded on the column in mobile phase A (0.1% (v/v) FA in water) and separated with a linear gradient of 3−35% mobile phase B (0.1% (v/v) FA in acetonitrile) at a flow rate of 300 nL/min over varying gradient lengths (30−240 min). The instruments were operated in a data-dependent mode and controlled by Xcalibur software (Thermo Fisher Scientific).

For the LTQ-Orbitrap Velos, a survey scan was acquired over the range *m*/*z* 350−2000 at a mass resolution of 30,000 (fwhm at *m*/*z* 400) and the top 20 most intense precursor ions were subjected to CID (normalized collision energy = 35, MS target value = 1.00E6, MS/MS target value = 1.00E4, and maximum ion fill time = 500 and 100 ms for MS1 and MS/MS scans, respectively).

For the Q Exactive HF analyses, an instrument method based on that described by Scheltema and colleagues[39] was used. A survey scan was acquired over the range *m*/*z* 350−2000 at a mass resolution of 60,000 (fwhm at *m*/*z* 200), and the top 18 most intense precursor ions were subjected to HCD (normalized collision energy = 28, MS target value = 3.00E6, MS/MS target value = 1.00E5, and maximum ion fill time = 100 and 45 ms for MS1 and MS/MS scans. respectively). The precursor ion isolation window in the quadrupole was set to 1.2 *m*/*z* units. Dynamic exclusion time was set to 20 s in all experiments.

Raw instrument data were processed with the MaxQuant (v. 1.5.1.0) software suite.[40] Peptide searches were performed against the UniProt *S. cerevisiae* canonical + isoform protein database (accessed on May 9, 2015; 6721 entries) and a database containing 262 common laboratory contaminants using the integrated Andromeda search engine.[41] Raw data from UPS and *E. coli* spike-in experiments were searched against the same *S. cerevisiae* protein database and UPS fasta file downloaded from the Sigma-Aldrich Web site (http://www.sigmaaldrich.com/content/dam/sigma-aldrich/life-science/proteomics-and-protein/ups1-ups2-sequences.fasta) or an additional *E. coli* strain K12 database (UniProt *E. coli* MG1655 reference proteome, 4303 entries), as appropriate. Search parameters were as follows: peptide false discovery rate (FDR) = 1%, protein FDR = 1%, precursor ion mass tolerance = 5 ppm (Velos) and 4.5 ppm (QEx-HF), product ion mass tolerance = 0.5 Da and 20 ppm, respectively. Carbamidomethylation of cysteine residues was set as a fixed modification, whereas protein *N*-terminal acetylation and methionine oxidation were set as variable modifications. A maximum of two missed cleavages per peptide were allowed, and matching between runs was enabled. All other MaxQuant parameters were left as default.

Individual peptide intensities were extracted by MaxQuant as a value at maximum of the MS1 peptide peak with the intensity threshold set to 500. This corresponds to the raw intensity value listed in the MaxQuant "peptides.txt" file.

Bioinformatics Data Analysis and Visualization

The MaxQuant output file containing individual peptide identifications and intensity information (peptides.txt) was further processed so that only unique peptides identified in at least three biological replicates were retained. For the purposes of this study, raw un-normalized peptide intensity values from the MaxQuant output were aggregated over the four replicates acquired for all yeast samples. This was done largely as a convenience to provide real values for all peptide instances even when values are very low or missing in one of the replicates.

For calculations of peptide intrinsic detectability or "flyability" factors (F-factors), absolute protein abundance values, in copies per cell, were taken from the CoPY project.[35,36,38] To retain only the highest quality measurements and reduce any ambiguity, this set was filtered so that only unique protein quantifications based on two SRM-measured peptides were included. Additionally, the individual peptide quantification values had to agree within 2-fold or better, i.e., the peptide A/peptide B ratio was two or less. This resulted in 349 yeast protein abundances expressed in copies per cell with a robust CV less than 16% calculated across four biological replicates.

Peptide F-factors were then calculated for all peptides identified in the label-free experiments that were contained in one of these 349 SRM-quantified proteins. F-factors were expressed as a ratio of raw MaxQuant-derived peptide intensity to copies per cell of their parent protein (or fmol concentration in the case of UPS standards) from which that peptide was derived, according to eq 1

$$\text{F factor}_i = (\text{intensity}_{ji}/\text{abundance}_j) \qquad (1)$$

where intensity$_{ji}$ is the intensity of peptide *i* from protein *j*.

All files necessary to reproduce the main results presented here are included as Supporting Information: Supplementary Tables 1A and 1B contain the raw peptide intensities from MQ from the UPS1 and UPS2 experiments. Supplementary Table 2 contains calculated F-factors for the UPS peptides. Supplementary Table 3 contains absolute abundance values of the 349 yeast proteins used for peptide F-factor calculations taken from ref 36. Supplementary Tables 4−7 contain MQ peptides.txt output for the primary experimental data sets used in this study, which are summarized in Table 1. The raw mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD003472.

DOI: 10.1021/acs.jproteome.6b00048
J. Proteome Res. XXXX, XXX, XXX−XXX

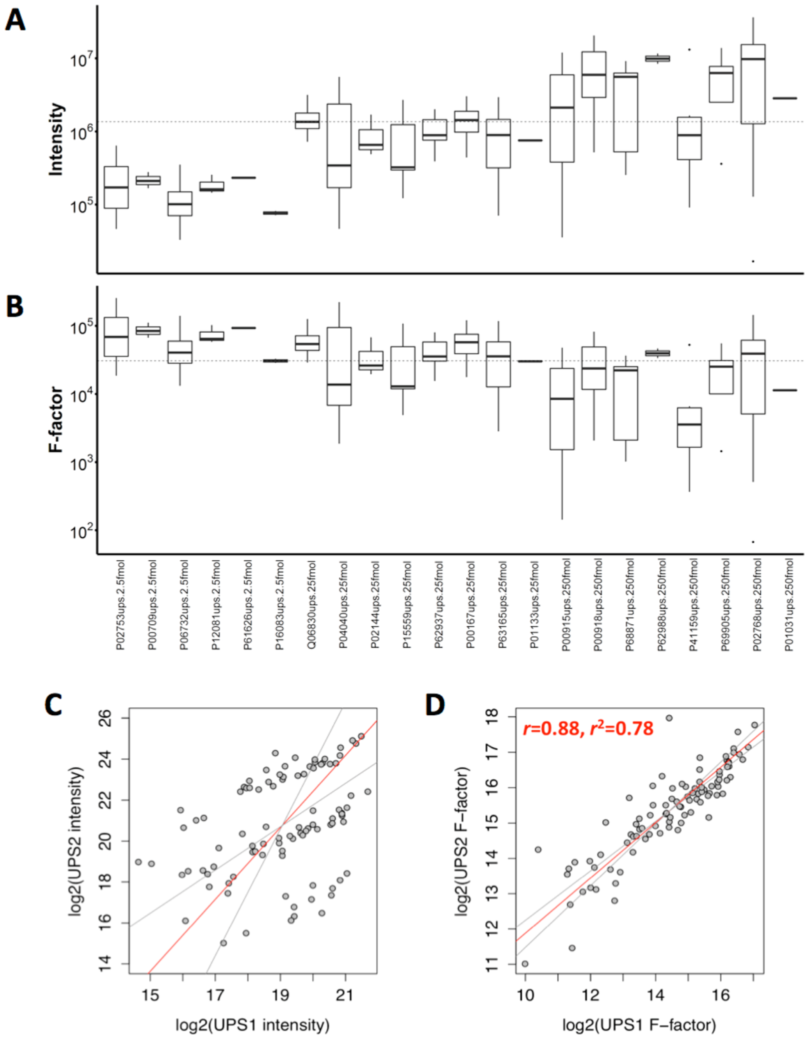


Figure 1. Peptide intensity and F-factor distributions from the experiments with the UPS standards. (A) The distributions of peptide intensity values from the UPS2 protein mix, detected on the LTQ-Orbitrap Velos instrument in a yeast background, are plotted as boxplots. Each boxplot corresponds to one UPS protein. Additionally, the proteins are grouped by their three concentrations, lowest to highest from left to right. The median peptide intensity is shown as a dotted line. (B) Corresponding plot for the F-factor distributions. The median F-factor is shown as a dotted line. (C, D) Peptide intensities and F-factors between UPS1 and UPS2. In C, UPS1 raw intensities are plotted against UPS2 intensities. The three UPS2 concentration levels appear as "lines" in the plot. In D, a corresponding scatterplot is displayed for the matched F-factors. It is visible that the correlation between UPS1 and UPS2 greatly increases, demonstrating that the abundance differences visible in C have effectively been removed by F-factor normalization. The red lines in C and D correspond to the regression slopes, and the gray lines are the ±95% confidence limits.

Definition of F-Factor-Based Peptide Classes

To investigate physicochemical properties of the peptide data sets, we partitioned the peptides into detectability classes based on their F-factors: "strong" flyers, "weak" flyers, and "nonflyers". For each of the four label-free experiments, the strong flyers data set was defined as the top 20% across the entire distribution and weak flyers as the bottom 20%. For comparative purposes, the negative data set of nonflyers was created in a similar way to that defined previously.[28] An in silico digest of the 349 yeast proteins was cross-referenced with all four label-free experiments and the stable isotope dilution (SIL)-SRM-MS data set, limiting potential peptides to those within the same peptide length limits as identified in the label-free experiments and with two or fewer missed cleavages. Peptides observed in any of the four experiments were excluded from the "nonflyer" set as well as any peptides that contained a subsequence that had been observed. In addition, using the MC:pred missed cleavage prediction tool,[42] any missed cleavage peptides with an internal missed cleavage score of >0.6 were excluded, reasoning they would unlikely to be observed as limit peptides. This yielded a filtered "nonflyers" data set of 10,577 peptides from 349 proteins.

Physicochemical Properties with AAIndex and Feature Selection

Physicochemical properties were assigned to individual peptides using an in-house version of the AAIndex resource (containing 544 different physicochemical properties). Both the mean and summed feature values from the individual residues in each sequence were used. Further features were added, including the estimated isoelectric point and amino acid composition, resulting overall in the computation of 1180 features. Owing to the large number of properties, a feature selection process was implemented to identify the most discriminating between the strong and weak flyers. Following previous work,[28] the Kullback–Leibler (KL) distance (also referred to as information gain or relative entropy) was

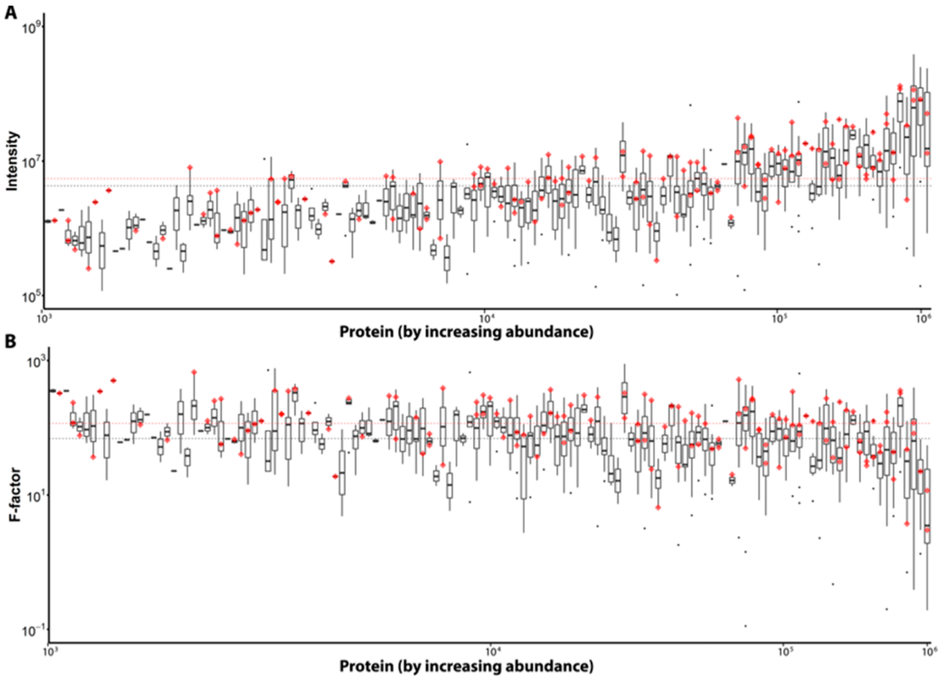


Figure 2. Peptide intensity and F-factor distributions from the standard yeast sample LC−MS/MS experiments. (A) The aggregated peptide-level intensity values from the LTQ-Orbitrap Velos 50 min experiment are plotted as boxplots by protein for those proteins for which copies per cell were determined by SIL-SRM-MS (as described in the Materials and Methods). The proteins are ordered by absolute abundance along the *x*-axis. The individual peptide intensities selected for the QconCAT SIL-SRM-MS study (from which absolute protein abundances were derived) are shown as red diamonds. (B) Corresponding plot for the F-factor distributions. Matched F-factors of the QconCAT SIL-SRM-MS peptides are again shown as red diamonds. In both A and B, the median value for all peptides is shown as a dashed black line, and the median for matched QconCAT SIL-SRM-MS peptides is shown as a dashed red line, which is the upper of the two lines in both panels.

calculated for every feature between the probability distributions of the strong and weak flyer data sets; larger KL distances equate to superior discrimination. Features with KL > 0.7 were then ranked and, descending down the list, features with absolute Pearson correlation coefficients to any previously retained feature >0.7 were excluded. This eliminated uninteresting features that closely matched any other to eliminate redundancy.

Methods for Peptide Hydrophobicity and Coelution Effects

In considering peptide hydrophobicity and coelution effects on ionization efficiency/detectability, only peptides identified in at least 3 out of 4 replicates were included; their F-factors were calculated as the median of replicate values. Also, all miscleaved peptides were removed, and only peptides eluting during the gradient (3−35% solvent B) were considered. These additional filtering steps ensured outliers in the data were removed.

For the coelution analysis, theoretical tryptic digests of the proteomes of *S. cerevisiae* and *E. coli* were computed using Protein Digestion Simulator (http://omics.pnl.gov/software/ProteinDigestionSimulator.php) with the following parameters: peptide fragment mass of 400−6000 Da, minimum peptide length of 7 residues, fully tryptic peptides (KR not P), and no missed cleavages. Retention times were predicted using SSRCalc[43] (see Supplementary Methods). Peptides were classified as coeluting if another peptide was detected within 25 s in the experimental runs, in any replicate, or predicted to elute within 25 s of another for the theoretical proteomes.

RESULTS AND DISCUSSION

F-Factors: a Peptide Ionization Response Metric

To evaluate F-factors as a metric of peptide ionization response and/or detectability, we first considered the peptide ion intensities assigned to UPS proteins,[44] a popular benchmarking standard, spiked into our standard yeast lysate as a background. In this experiment, because all UPS proteins are either at a single known concentration (UPS1: 25 fmol) or six different known concentrations (UPS2: 0.0025, 0.025, 0.25, 2.5, 25, or 250 fmol), F-factors can be directly assigned by normalizing the median peptide ion intensity across four replicates by the known concentration. Figure 1A shows boxplots of peptide ion intensities reported by MaxQuant following identification by Andromeda[41] using a 1% FDR for the range of detected UPS2 proteins analyzed on the LTQ-Orbitrap Velos. A total of 22 UPS2 proteins from the concentration range 2.5−250 fmol were detected, whereas all 48 proteins were detected in case of the UPS1 standard. Supplementary Table 2 contains all peptide identifications, median intensity, and calculated F-factors from the UPS experiments. From the UPS2 experiment, detected peptide ion intensities are clustered into three broad ranges corresponding to the three spike-in concentrations presented in increasing order from left to right on the plot. The equivalent F-factor plot is shown below in Figure 1B, where the normalization process has removed much of the variation attributable to the three UPS protein levels in the analyte. By comparing the matched signals between proteins in the UPS1 and UPS2 standards directly, panels C and D in Figure 1 show the correspondence in the data. The raw intensities are not universally correlated between the two experiments, and the three UPS2 spike-in concentrations detected are apparent as three "lines" across the intensity plot when compared to UPS1 median intensities (collected at a single concentration). In contrast, the F-factors agree very well between the UPS1 and UPS2 runs as shown in Figure 1D. The differences are representative of the technical variation in the experiment, much of this likely due to the differences in chromatographic performance in the context of a complex yeast digest background. Globally, we observe UPS peptide F-factors spanning a dynamic range of around 3-fold, though this is closer to 2-fold on a per-protein basis as illustrated in Figure 1B.

The UPS experiment considers a well-characterized benchmarking standard, but not the F-factors of a native proteome over a large dynamic range. We therefore collected four primary data sets obtained from nESI LC−MS/MS experimental analysis of yeast whole cell lysate obtained from chemostat cell culture. The lysate was identical to that used previously for the parallel, QconCAT-based, large-scale absolute protein quantification study.[35,36,38] This yeast sample was chosen because it is a well-characterized standard in our laboratory for which high quality absolute abundance values have been determined by SIL-SRM-MS.[36] The resulting numbers of peptide identifications from the label-free LC−MS/MS experiments are summarized in Table 1, detailing the instrument employed as well as the total number of peptide identifications achieved (1% FDR) using the MaxQuant search engine Andromeda.[41] For peptide F-factors to be calculated, MaxQuant-derived peptide intensity values were normalized by the equivalent protein copies-per-cell value. The ranges of peptide ion intensities assigned by MaxQuant for individual proteins, for a single label-free experiment (LTQ Orbitrap Velos 50 min gradient), are shown in Figure 2A; the boxplots are ordered according to protein abundance as determined by the SRM data (increasing from left to right on the *x*-axis). For clarity, we chose to display the label-free experiment with the fewest proteins identified, although identical trends were observed across all runs (Supplementary Figure 1). As expected, in general, peptide ion intensity increases with higher protein abundance, and the median intensity for each protein correlates well with absolute protein abundance (Spearman correlation = 0.78−0.86 across the four label-free runs). It is apparent from the distributions that some peptides give rise to much higher signal intensities than others despite being in a similar concentration range. Typically, as observed for UPS proteins, peptide intensities from the same protein span two, or occasionally three, orders of magnitude. Furthermore, many peptides derived from the lower abundance proteins give stronger signals than peptides present at much higher concentrations. For example, peptide LVIPDILTR from the protein YDR341C at ∼3,000 copies per cell produces a summed ion count from the four replicates of 1.3×10^6, whereas LQQTAFDK from protein YKL056C at ∼120,000 copies per cell has a lower summed ion count of 0.9×10^6.

Another notable feature across all data sets is the relative position of the intensities of the surrogate Q-peptides selected for SRM analysis, shown as red diamonds on the individual boxplots. These peptides were selected on the basis of their proteotypic properties,[28] low predicted miscleave propensity,[42] and heuristic rules intended to eliminate poorly ionizing peptides and select good quantotypic candidates for targeted analysis.[35] As can be seen, this was generally highly successful as most red diamonds are in the upper quartile of the peptide intensity distributions and frequently at or close to the top, indicating they are readily ionized and detected. Indeed, the median value of the matched Q-peptide intensities/F-factors (red dashed line, Figure 2A,B) exceeds the equivalent value calculated for all peptides. In total, 137 of the 698 Q-peptides used for the SIL-SRM-MS studies were observed in the 50 min LTQ-Orbitrap Velos run, whereas 347, 402, and 485 were observed in the 240 min Velos experiment and 50 and 120 min gradient on the Q Exactive HF experiments, respectively. Clearly, there were also cases where a Q-peptide for a given protein had a poor F-factor or was not observed at all in the label-free experiment; this is likely due in part to the stochastic sampling nature of a data-dependent analysis (DDA) experiment failing to select them for tandem MS and highlights the continued advantages of a targeted approach. During the selection of quantotypic peptides for QconCAT design, it was sometimes the case that no strong candidate tryptic peptides passed all of the rigorous selection filters, leaving limited choices for selection of an internal reference peptide; equally, we recognize that on occasion some peptides were simply poor selections that with hindsight might not have been chosen. In any case, the use of an identical heavy-labeled standard in the SRM experiments ensures that even poor flyers are quantified accurately.

Normalizing the peptide intensities to generate F-factors removes the trend for increasing peptide intensity with protein abundance, as illustrated in Figure 2B, for the same LTQ-Orbitrap Velos experiment as Figure 2A. Peptides with large F-factors give rise to disproportionally greater ion current than would be expected based on their amount in the sample. Equally, low F-factor values suggest a peptide "underperforms", resulting in disproportionately fewer ions reaching the detector. Despite some anomalies at low and high protein abundance, it can be seen that most F-factor distributions are broadly centered about a common median value. Although peptide intensities are positively correlated with protein abundance (Kendall's nonparametric tau-b coefficient = 0.47) their F-factors should not be, which is indeed the case (Kendall's tau-b = −0.21). Therefore, the F-factor represents a useful intrinsic measure of peptide detectability and effective ionization efficiency that removes much of the confounding abundance bias,[11,37] notwithstanding additional caveats that we discuss further below.

One notable feature, visible in Figure 2B, is the apparent enrichment for low F-factors in high-abundance proteins. This can be rationalized by considering that at very high protein abundance, the number of peptide ions reaching the detector at any given time may exceed the limits of the system (e.g., may saturate an electron multiplier detector or lead to a space charge effect inside an ion trap). In the more recent orbitrap instruments, the space charge effect is mitigated to some extent by the use of automatic gain control, which attempts to optimally adjust the number of ions filling the mass analyzer at each scan. Nevertheless, regardless of the source of the effect, the signal of highly abundant ions could still be outside of the linear dynamic range and outside of the upper limit of quantification (ULOQ). Lower raw peptide ion intensities would thus be recorded in the label-free experiments, and consequently, F-factors would be underestimated. The high dynamic range (over 5 orders of magnitude) in a eukaryotic proteome[45] means that MS1 quantification over the whole protein abundance range is known to present an issue in these types of shotgun experiments. Although we cannot exclude the possibility that some QconCAT SRM-based values may be

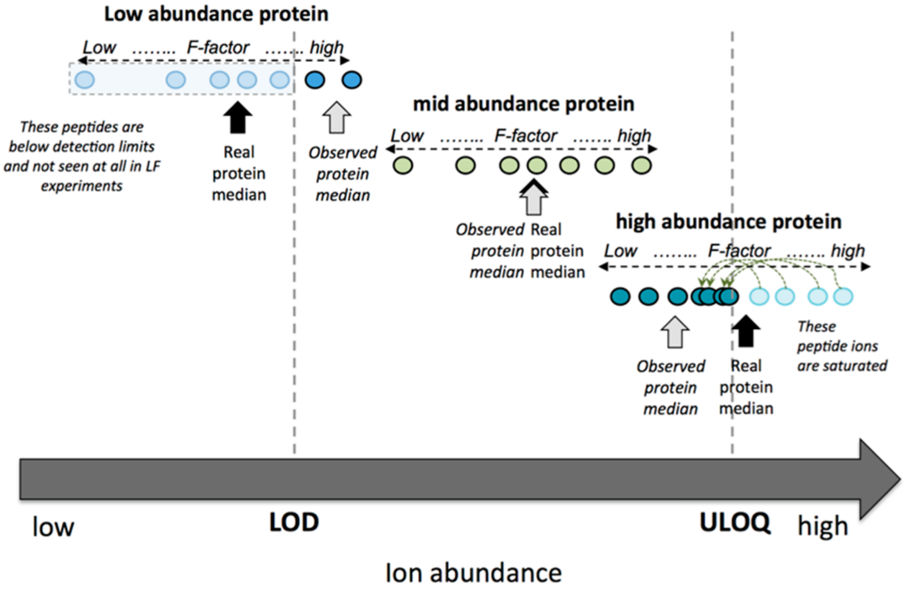

Low abundance protein
Low …….. F-factor ……. high
These peptides are below detection limits and not seen at all in LF experiments
Real protein median
Observed protein median
mid abundance protein
Low …….. F-factor ……. high
Observed Real protein protein median median
high abundance protein
Low …….. F-factor ……. high
These peptide ions are saturated
Observed protein median
Real protein median
low
LOD
ULOQ
high
Ion abundance

Figure 3. Schematic illustrating peptide detection effects for low- and high-abundance proteins. Three theoretical proteins are shown, each with a range of peptides displaying different peptide ionization properties leading to different F-factors. At low protein abundance, the poorer ionizing peptides fall below the effective detection limit (LOD) of the instrument, whereas at high protein abundance, saturation effects lower the observed ion signal detected. In both cases, the observed protein median F-factor will be altered from its true value, leading to the effects visible in Figure 2.

overestimates of the true protein concentration leading to underestimated F-factors, this seems unlikely; the SIL-SRM-MS quantification used two closely correlated internal standards, in each case selecting the closest of four spiked-in concentrations of stable-isotope standard to determine the absolute level.[35]

This phenomenon, and the counter-effect at low abundance, is illustrated schematically in Figure 3. At low protein abundance, fewer molecules will be present, and only peptides that are readily ionized, transmitted, and fragmented are likely to be observable and detected. This would lead to a bias toward high F-factor peptides. Equally, the MS target value and maximum ion fill time settings used to accumulate ions in trap-based hybrid mass spectrometers could also lead to an overestimation of the inherent F-values for some peptide ions, for example, for peptides eluting at times where the total ion current is below the predefined target value and ions are accumulated until that target is reached. In either case, both of these rationalizations would lead to apparent overestimates for the protein-based median F-factor for low abundance proteins, which is observed for all of the data sets considered here.

In support of these observations, we note that peptides from the top 20% of all F-factors collated across one experiment are disproportionately represented by proteins of low abundance (Supplementary Figure 2C) where the fraction of peptides detected is also low (Supplementary Figure 2D).

Conservation of F-Factors and Peptide Detectability

As shown by the kernel density plots in Figure 4A, there is a broad distribution of F-factor values on the two LC−MS platforms employed in this study, reflecting the wide range of effective peptide ionization efficiencies in a typical proteomics experiment. The distribution of F-factor values appears to be generally well conserved in their overall shape and properties between platforms and differing chromatography regimes, although the F-factor values are higher for the two Q Exactive HF data sets than from the LTQ-Orbitrap Velos. This is likely an instrument-specific factor given the differences in ion transmission between the platforms and the fact that the two instruments record ion current on a different numeric scale. However, the paired distributions of F-factors from the same mass spectrometers collected from different LC gradient times are similar with medians of 69 and 50 for the 50 and 240 min gradients on the LTQ-Orbitrap Velos and 10.8 × 10^3 and 9.0 × 10^3 for the 60 and 120 min gradients on the Q Exactive HF, respectively. The slightly lower values observed for the longer gradients on both platforms reflect the deeper sampling of the proteome: as more "weak" flyers are detected, the overall median goes down.

To compare the individual F-factors between runs, we considered the overlapping groups of peptides identified in paired experiments. Figure 4B shows the correlation between peptide matched F-factors, which is strongest between runs conducted on the same MS platform and closely mirrors the good correspondence observed for the UPS standard proteins shown in Figure 1. For example, the Pearson correlation coefficient between the two Velos runs is 0.78 (r^2 = 0.61) and 0.88 (r^2 = 0.77) between the Q Exactive HF runs. As might be expected, the corresponding correlations across instruments are weaker. This data shows that there is a reduced but reasonable correspondence between the intrinsic detectability responses across mass spectrometric platforms using nano-ESI.

Figure 5A shows the overlap between F-factors determined in the four shotgun experiments for 796 peptides that were universally detected from the 349 targeted protein set. These correspond to the expected "low hanging fruit", i.e., peptides from more abundant proteins that are also readily ionized, fragmented, and detected across all platforms. Indeed, the median abundance of parent proteins associated with the overlap set is 57,000 copies per cell. The additional peptides identified on the longer Velos gradient, for example, are from lower abundance proteins with median abundance of only 16,000 copies per cell. This is of course expected and highlights large gains in proteome coverage as a result of improved capability of the experimental strategy employed: longer chromatographic gradients in the case of Velos and higher

DOI: 10.1021/acs.jproteome.6b00048
J. Proteome Res. XXXX, XXX, XXX−XXX

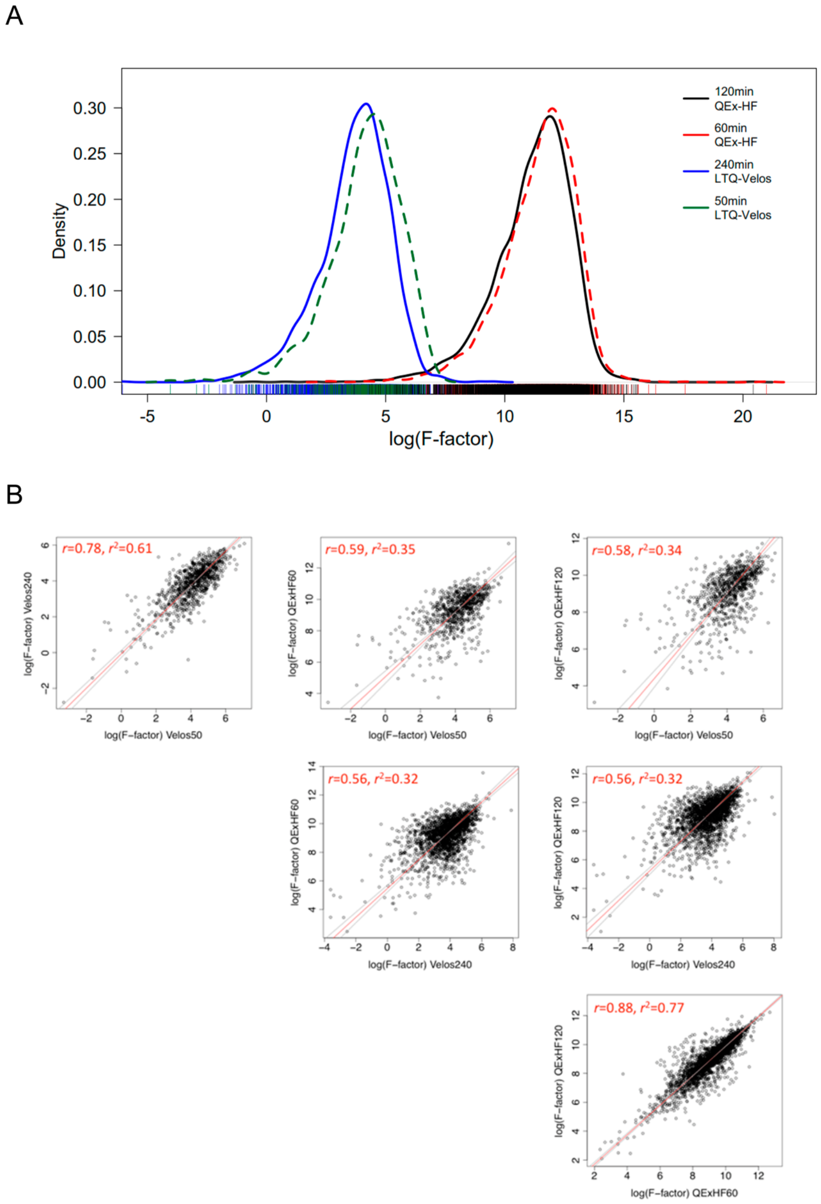


Figure 4. Effects of chromatographic conditions and instrument platforms on effective peptide ionization. (A) F-factor distributions for the four experiments used in this study, illustrating a broad range of peptide ionization efficiencies maintained across chromatographic conditions but globally shifted between instrument platforms. (B) Scatterplots comparing F-factors between paired shotgun LC−MS/MS experiments for the four different runs are shown. A fair correlation between experiments is observed, most significantly for data originating from the same instrument. The changes of individual peptide F-factors likely reflect varying matrix effects between gradients and ion source and transmission efficiency between instruments.

MS2 resolution and greater achievable acquisition speed on the Q Exactive HF. To investigate further, we compared the F-factor distributions of the common 796 peptides present on all platforms to those unique to a particular run and gradient. As can be seen in Figure 5B, there is a significant difference in F-factor distributions (Wilcoxon test, $p < 2 \times 10^{-16}$) between the common peptide F-factors and those unique to each of the four platforms, which have universally lower detectabilities. We ascribe this again to the deeper sampling of the proteome generally possible via longer gradients or on an alternative, theoretically more sensitive, instrument that is able to support detection of peptides that are outcompeted for ionization or otherwise below the detection level. Although these observations are unsurprising, they provide further evidence that F-factors provide information on the detectability properties one would expect, representing a good intrinsic metric.

F-Factors and Peptide Physicochemical Properties

To classify peptides by their relative detectability and effective ionization efficiency, we split them into two groups: "strong flyers" and "weak flyers". Strong flyers were the top 20% of peptides with highest F-factor values, and weak flyers were the bottom 20%. Additionally, for comparative purposes, we defined a negative "nonflyers" peptide group. The "nonflyers" set consisted of tryptic peptides that could potentially be observed but were not detected in any of the shotgun proteomics experiments conducted in this study. Next, we examined the physicochemical properties of the peptides in these three groups to determine which of the properties are important determinants of ionization efficiency. The rationale behind this exercise was that F-factors remove the confounding signal arising from abundance and therefore represent a better

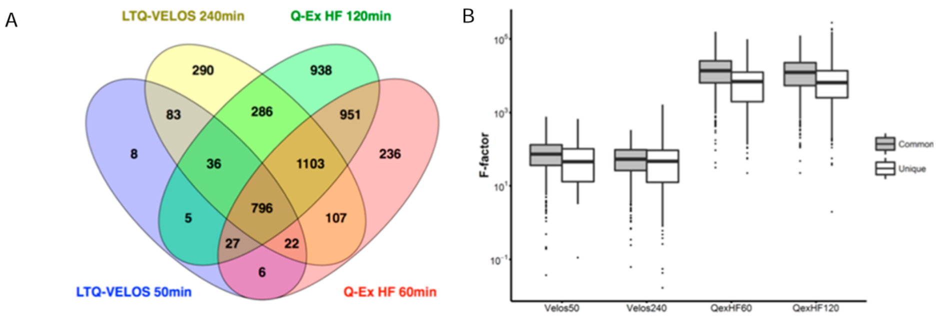


Figure 5. Correspondence between F-factors in the four LC−MS/MS experiments. (A) Overlap between high quality peptide identifications observed in the four shotgun experiments on the yeast proteome shown as a Venn diagram, restricted to unique peptides (FDR < 0.01) mapping to one of the 349 SIL-SRM-MS quantified proteins. (B) Boxplots showing the distribution of F-factors for the 796 common peptide set compared to those determined uniquely by individual experiments. There is a significant difference between the paired distributions for all but the Velos 50 min data (Wilcoxon rank; $p < 0.02$ Velos 240 min, $p < 2 \times 10^{-16}$ Q Exactive HF runs).

Table 2. Physiochemical Properties Discriminating F-Factor Classes

parameter name	AAIndex or other description	KL distance	correlation	type
sum of methionine residues	total count of M	2.47	negative	other
sum: VENT840101	bitterness[a]	2.45	positive	hydrophobicity
sum of neutral residues	total count of all bar (D,E,K,R)	1.92	positive	charge
mean of basic residues	length normalized count of (H,R,K)	1.75	negative	charge
sum: CHAM830104	number of atoms in the side chain labeled 2 + 1[b]	1.17	positive	structural
pI (isoelectric point)	estimated isoelectric focusing point	1.14	negative	charge
sum of basic residues	total count of (H,R,K)	1.12	negative	charge
mean: VENT840101	bitterness[a]	0.88	positive	hydrophobicity
mean of charged residues	length normalized count of (D,E,K,R)	0.88	negative	charge
sum: nosheet	enrichment in beta-sheet forming amino acids (I,F,T,W,Y,V)	0.84	positive	structural
mean: CHAM830108	parameter of charge transfer donor capability[b]	0.81	negative	charge
mean: surface	length normalized count of typical accessible amino acids (R,N,D,E,Q,G,H,K,P,S,T,Y)	0.78	negative	hydrophobicity
sum of arginine residues	total count of R	0.77	negative	charge
mean number of aromatic residues	total count of (H,F,W,Y)	0.76	negative	hydrophobicity

[a] Venanzi *J. Theor. Biol.* 1984; 111, 447−450. [b] Charton and Charton *J. Theor. Biol.* 1983; 102, 121−134.

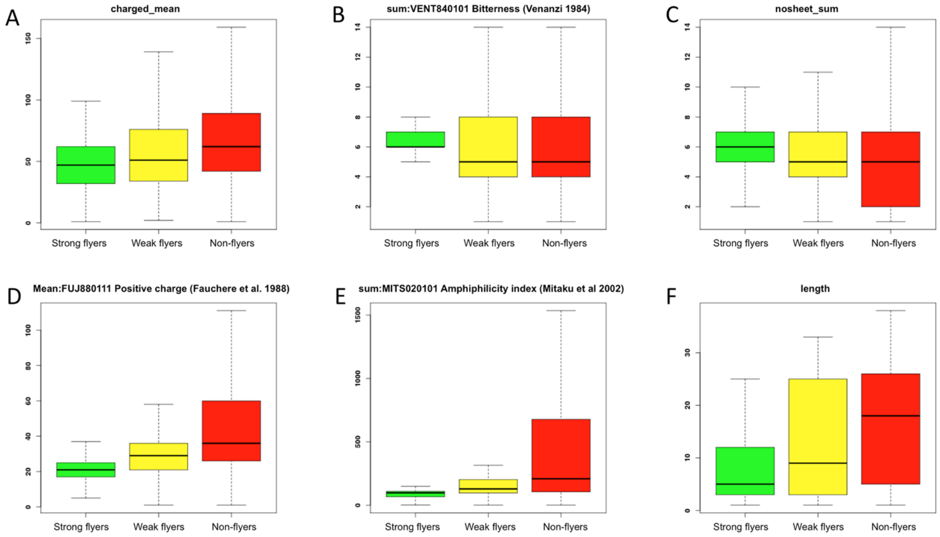


Figure 6. Selected physicochemical properties of strong, weak, and nonflying peptides. The plots highlight differences in physiochemical feature distributions between different peptide flyability classes with the corresponding calculated AAIndex value[46] or physical parameter on the vertical axis of each panel. The displayed physiochemical properties are the ones showing the greatest differences between the classes and are therefore likely to affect peptide ionization the most. (A−C) Significant features observed on the QExactive HF, and (D−F) on the LTQ Orbitrap. AAIndex codes or features names are listed above each panel.

DOI: 10.1021/acs.jproteome.6b00048

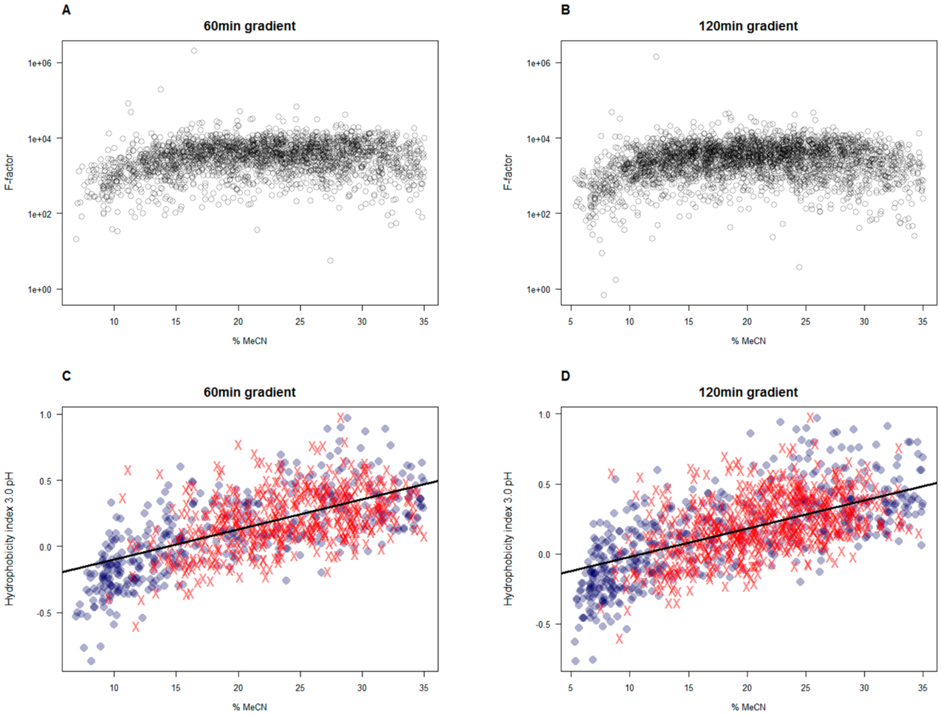


Figure 7. F-factor and peptide hydrophobicity across elution profiles. (A and B) Variation in F-factor and (C and D) calculated peptide hydrophobicity observed over the analytical gradients used in the Q Exactive HF experiments of 60 and 120 min, respectively. Profiles are shown as the concentration of acetonitrile increases, shown on the *x*-axis. Low and high F-factor peptides are shown as blue circles and red x's, respectively.

measure than the raw intensities or binary classifications of peptide proteotypic presence/absence as used in previous studies.[28−32] To compute the physicochemical properties (or features as they would be referred to in the machine learning field), we used the AAIndex,[46] a public resource that stores residue-level parameter sets for hundreds of amino acid properties published in the literature.

Following the feature selection procedure, as described in the Materials and Methods section and in more detail in ref 28, we were able to find a number of features that distinguish between strong and weak F-factor groups and hence are likely determinants of peptide detectability/ionization efficiency (Table 2). The most informative were generally related to the number of basic residues, peptide length, and hydrophobicity. Many of these properties have been documented in the literature as contributing toward peptide ionization and fragmentation, for example, in refs 28, 29, 31, 47, and 48. Our analysis further supports those studies and their findings. However, the important distinction here is that these properties were obtained based on peptide groups defined by F-factors and thus normalized for abundance, as opposed to peptide groups defined simply by presence/absence or other metrics confounded by abundance. Figure 6 shows some exemplar distributions of feature scores for differing peptide categories from the Q Exactive HF (Figure 6A−C) and LTQ-Orbitrap Velos (Figure 6D−F) data sets. Similar distributions for each related feature were also observed in the other two experiments not illustrated.

For example, Figure 6D highlights the effect of the number of basic residues (which will be positively charged at low pH) on F-factor values. This feature, although frequently identified as important for ionization by proteotypic peptide predictors, has also been coupled to fragmentation efficiency. This is generally rationalized by the mobile proton model[49,50] and the hypothesis that the efficiency of collision-induced dissociation is enhanced under conditions of a mobile proton environment (where the number of ionizing protons is larger than the number of basic residues).[51] Indeed, a negative correlation between the number of basic residues and flyer type is observed here (Figure 6A, Table 2). The effect of net charge on peptide flyability is further supported by considering peptide length, where shorter peptides appear to have improved detectability (Figure 6F). On this basis, tryptic peptides with lower average number of basic residues and shorter length will thus be expected to generate a more complete set of fragments for identification by search engine algorithms;[52] indeed, the peptides classed as strong flyers (higher F-factor) have on average a more complete ion series, as illustrated in Supplementary Figure 3.

Peptide Hydrophobicity and Detectability

Hydrophobicity is widely reported as a determinant of ion detectability, which is also confirmed to be the case here (Table 2), and is often used as a principal feature in machine learning approaches to select peptide surrogates for targeted proteomics.[28,29] Typically, it is positively correlated with detectability, although a recent report suggests this might not always be true, particularly in case of MS2-based quantification.[48]

More generally, it is unclear whether peptide hydrophobicity itself is an intrinsically important physiochemical property promoting enhanced ionization or is incidentally associated for other mechanistic reasons. Specifically, hydrophobicity may be linked to improved evaporation from droplets during ESI because hydrophobic peptides are more likely to be located toward the droplet surface.[22,25] Additionally, peptides eluting later in the reversed phase gradient will be more hydrophobic and, hence, are likely to ionize better simply due to improved desolvation at the higher organic solvent concentrations required for their elution.[53]

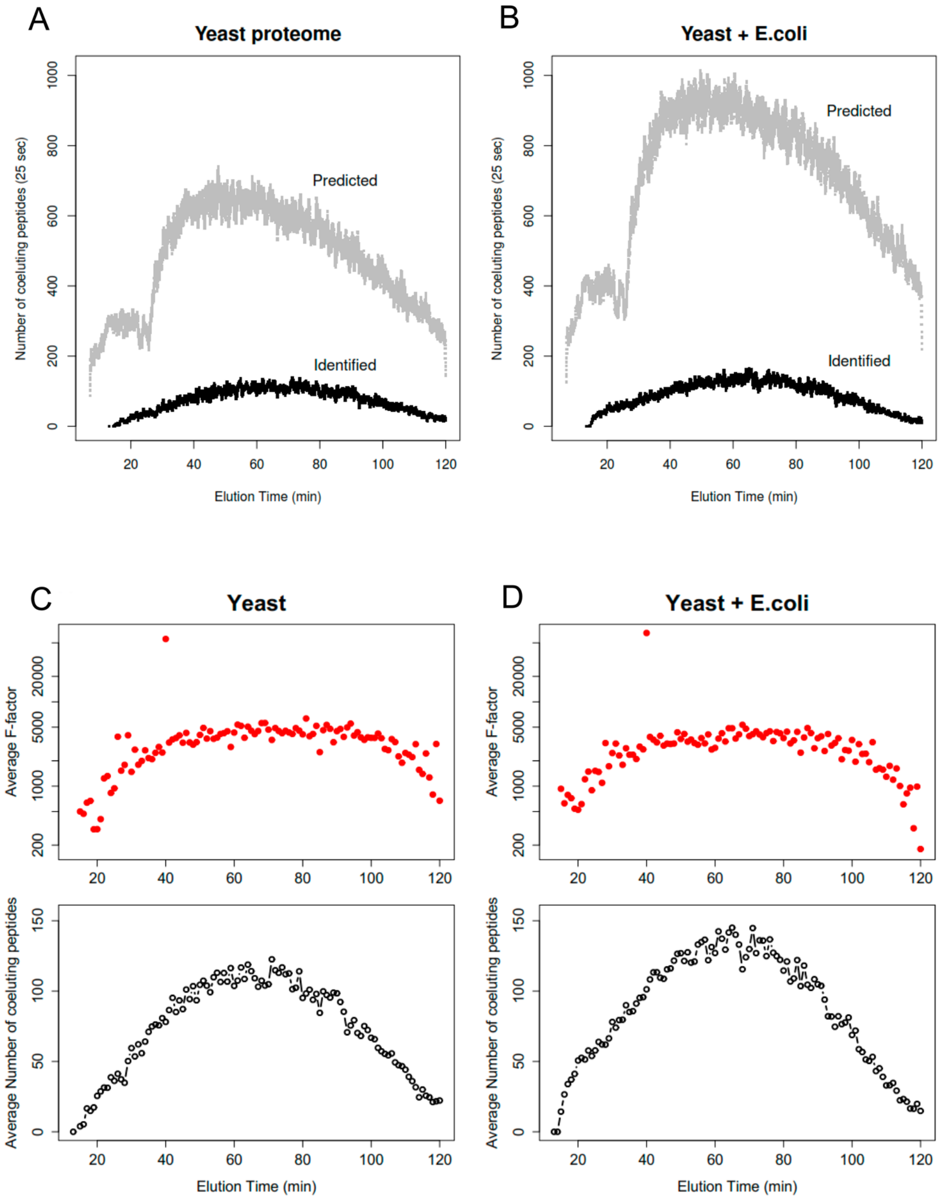


Figure 8. Coelution profiles for yeast and yeast + *E. coli* samples during a 120 min gradient. (A, B) Retention time is plotted against the number of peptides within a 25 s chromatographic window centered on a given peptide. Each point therefore represents a peptide that is eluting at a given time (displayed on the *x*-axis) and is surrounded by other coeluting peptides (a number of which are displayed on the *y*-axis). The black traces show the identified coeluting peptides in each run, and the gray traces show the predicted coeluting peptides (following assumptions outlined in the text). (C, D) Peptide F-factor profiles and coelution properties. Average F-factor values are shown as a function of elution time for the (C) Yeast120 and (D) YeastEcoli120 samples. In the panel directly below, the corresponding average number of coeluting peptides (black trace) as a function of time is displayed.

To discriminate these possibilities, we examined whether detectability increases linearly with higher concentrations of organic solvent. For this exercise, only peptides eluting between 3 and 35% MeCN were considered. Panels A and B in Figure 7 show that there is no strong linear relationship between F-factor and % acetonitrile required for peptide elution in either the 60 or 120 min chromatographic gradient. Instead, there appears to be a weak nonlinear effect; ionization efficiency is reduced for peptides eluting at both low and high organic concentrations. Mechanistically, this could be explained by two opposing processes: at very low organic concentrations, where the surface tension of the droplet is high, ion desolvation is hampered; at higher organic levels, desolvation is enhanced but the increased gas phase basicity of acetonitrile makes it more likely to interact with protons and therefore reduces their availability to ionize peptides.

This weak, nonlinear effect is further emphasized in Figure 7C and D where the percentage of mobile phase B is plotted against peptide hydrophobicity index in acidic pH[54] for low and high F-factor peptides. As expected, the average peptide hydrophobicity increases across the gradient, but there is a wide range of individual hydrophobicity values at any one time. Additionally, a greater proportion of weak flyers is observed at both low and high organic solvent concentrations. Although simple peptide mixtures might display increasing signal response with increasing retention times or organic concentration (as observed for example by Cech et al.[55] and Osaka and Takayama[27]); in other cases, the opposite effect has been

DOI: 10.1021/acs.jproteome.6b00048

reported.[56] Our data shows that, in a complex mixture where thousands of peptides are present, there is no longer a simple linear dependency. Indeed, the data presented here shows that hydrophobicity is not a simple monotonic predictor of peptide detectability, suggesting instead that extreme hydrophobicity values disfavor ionization.

Peptide Coelution and Competition for Ionization

Chromatographic coelution happens when two or more analytes cannot be separated because the difference in their retention is smaller than the chromatographic resolution. In a nano-ESI−MS/MS proteomics experiment, coeluting peptides can be defined as those that are simultaneously sprayed into the instrument. The resulting competition for ionization is expected to have a significant effect on peptide ionization efficiency and consequently detectability. Peptides analyzed under different chromatographic regimes (as for example illustrated in Figure 4B) will experience altered signal suppression arising from changes in the number of coeluting peptides. Therefore, to investigate the effect of coelution on peptide detectability, we introduced competition for ionization without changing the chromatographic gradient by spiking in an *E. coli* proteome to the yeast sample. Data were then compared to the pure yeast sample under the same chromatographic conditions. The two data sets are referred to as Yeast120 and YeastEcoli120. On the basis of the average chromatographic peak width, we considered a peptide to be coeluting if its retention time (calculated as an average from four replicates) was within a 25 s time window centered on another peptide. Alternative time windows (20, 30, and 60 s) were also tested and yielded similar results (data not shown).

First, we considered the peptides that were successfully identified in each run. On average, there were 89 and 105 detected and identified peptides coeluting with any given peptide in the Yeast120 and YeastEcoli120 samples, respectively. The coelution profiles are not uniform along the gradient, and the largest number of coeluting peptides, over 130 for Yeast120 and 160 for YeastEcoli120, appear in the middle (illustrated as the black trace in Figure 8A for Yeast120 and 8B for YeastEcoli120). However, the real numbers of coeluting peptides are likely to be much higher (Figure 8, gray trace) because many peptides present in the sample are not selected for fragmentation and are thus not identified. A theoretical tryptic digest of the yeast proteome should contain over 140,000 unique peptides and yeast + *E. coli* proteomes over 220,000 (assuming only limit peptides with a minimum length of 7 residues). We used these peptide data sets to estimate the theoretical retention time profile for all tryptic peptides from *S. cerevisiae* and *E. coli* proteomes based on a simple linear model calibrated on the measured retention times in each sample (for description, see the Materials and Methods section).

The gray traces in Figure 8 show that, when considering the full tryptic proteome, an average of over 440 additional undetected peptides are likely coeluting with any given peptide in the Yeast120 sample (Figure 8A) and over 660 peptides in YeastEcoli120 (Figure 8B). This becomes proportionally greater when average peak widths are longer; for example, a 30 s elution window is predicted to have 600 and 900 more undetected coeluting peptides for Yeast120 and YeastEcoli120 samples, respectively. Consequently, it is obvious that only a fraction of all peptides present in the sample are identified. This simple analysis highlights the well-known limitations of peptide detection and identification that are in part dependent on MS speed and sensitivity.[57,58] Indeed, in the context of coelution, a computational study by Schliekelman and Liu[59] showed that the number of coeluting peptides is a significant factor determining the probability with which a peptide is detected.

Here, we can expand on these observations by considering the direct effect of the number of coeluting peptides on the detectability of a peptide as estimated by its F-factor. Panels C and D in Figure 8 show the average F-factor and number of coeluting peptides as a function of time in Yeast120 (Figure 8C) and YeastEcoli120 (Figure 8D). It can be seen that average detectability (quantified by F-factor) initially increases with time and drops off toward the end of the gradient. This pattern is matched by the number of coeluting peptides. When the number of coeluting peptides (at the start and end of the gradient) is small, the average F-factor is lower as those peptides with small F-factors (lower ionization efficiency) are detected. When competition for ionization is at its highest (in the middle of the gradient coincident with a much higher number of coeluting peptides), the peptides with high F-factors (better ionization efficiency) are preferentially detected. This observation suggests that, in regions where competition for ionization is weak, even peptides with poor ionization properties can be detected. In contrast, in regions of the LC gradient where competition for ionization is fierce, intrinsic peptide detectability must be greater to outcompete coeluting peptides and be detected. Furthermore, the average F-factor for all peptides decreases for the YeastEcoli sample (mean fold change = −0.27), which is consistent with a greater number of molecules competing for the same amount of charge in the ESI droplet.

■ CONCLUSIONS AND PERSPECTIVE

Here, we introduce the concept of peptide flyability factor (F-factor) in bottom-up proteomics experiments, defined as the ratio of peptide signal intensity to its parent protein absolute abundance. We demonstrate that this is a useful metric, which effectively removes protein abundance bias from the measured peptide intensity and enables a better understanding of intrinsic detectability and, by proxy, ionization efficiency in the LC-ESI−MS experiment. Although the two are closely related, ionization efficiency is a more fundamental property that describes the fraction of gas-phase ions generated from the total number of molecules introduced. Detectability, on the other hand, depends both on ionization efficiency as well as other factors, like ion transmission efficiency and detector response. These and other contributing factors, such as protease cleavage efficiency and post-translational modifications, make the precise determination of ionization efficiency a much harder task. In the present study, we focused on the intrinsic peptide detectability (quantified by F-factors) rather than determination of the ionization efficiency per se. We argue that this is a valuable approach because abundance represents the most significant confounding factor with protein abundances typically varying over 4 orders of magnitude in yeast[60] and as high as 10 orders in human plasma.[61] We calculated F-factors for thousands of peptides in the yeast proteome, demonstrating that peptides derived from the same protein, and presumably at (almost) identical concentrations, ionize and are detected in the mass spectrometer with markedly different efficiencies. This was apparent, for example, from the UPS data set acquired here (UPS spiked into the yeast lysate); the average range of peptide F-factors per protein was around one order, but the highest approached 3 orders of magnitude. In contrast, the total dynamic range of the experiment (i.e., considering intensities of all detected and identified peptides) was around 4 orders of magnitude.

It is also important to note that other factors can potentially further contribute to the large range of intensities, for example, missed-cleavages, post-translational modifications, or solubility issues. Although these factors were not explicitly considered here, we believe they are not the most confounding; for example, eliminating peptides predicted to be poorly digested (i.e., containing missed cleavages) has only a modest effect on F-factor distributions, although it does slightly increase the median F-factor and reduce the variance (Supplementary Figure S4).

The normalization of raw peptide intensities by their abundance establishes F-factors as an intrinsically quantitative property that represents a better measure of detectability and intrinsic ionization efficiency of a tryptic peptide than its raw intensity. Consequently, F-factors show similar distributions across different experimental conditions and indeed, excluding some scaling details, across instrument platforms typically spanning 2 or 3 orders of dynamic range. In addition, F-factors should enhance selection strategies for targeted proteomics because the confounding issue of protein abundance is removed. In support of this, we note that F-factors retain the expected trends between strong flyers, peptides that tend to ionize and be detected well, and weak flyers/nonflyers. For example, when comparing strong, weak, and nonflyers, the strong flyers have a lower ratio of basic residues (His, Lys, and Arg) to their length, which intriguingly follows trends observed for fragmentation via the mobile proton model theory.[49] Similarly, F-factors support a broader investigation of the underlying physicochemical properties that mediate peptide electrospray ionization and other confounding issues, such as coelution. However, many additional contributions to peptide detectability exist that have not been explicitly considered, for example, peptide charge state or potential post-translational modifications, both of which will alter peptide detectability and F-factor as well as reduce the amount of analyte present. Finally, because F-factors show conservation across some experimental platforms and protein abundance is similarly conserved across species boundaries,[62] the insights obtained from our analyses are likely to be widely applicable, for example, in selection or prediction of the best-suited peptides for SRM experiments as well as statistical modeling and calibration of label-free data for more accurate, absolute quantification of proteins in complex biological samples.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00048.

> Methods describing additional protocols for calculating predicted retention time and coelution properties; figures describing the complete peptide F-factor distributions for subsets of the 349 proteins observed on all instruments, shown as boxplots; additional peptide properties for F-factor classes, and detection by protein abundance; percentage of fragment ion series detected for peptide subclasses; and peptide signal intensity/F-factor distributions considering missed cleavages (PDF)
>
> Detailed proteins, peptide signal intensities, and F-factor values for all runs (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: Claire.Eyers@liverpool.ac.uk; phone: +44 (0)151 795 4424.
*E-mail: simon.hubbard@manchester.ac.uk; phone: +44 (0) 161 306 8930.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge support from the Biotechnology and Biological Sciences Research Council in the form of a studentship to A.F.J. as part of the Systems Biology Doctorial Training Centre, grant support to S.J.H. and C.E.E. in the form of a LoLa grant (BB/G009112/1), and additional support for C.L. (BB/L002817/1). S.W.H. is supported under grant BB/M025756/1. D.C.H.L. acknowledges Cancer Research UK for their support.

REFERENCES

(1) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **2013**, *49* (4), 583−90.

(2) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol. Cell. Proteomics* **2014**, *13* (1), 339−47.

(3) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **2012**, *11* (3), M111 013722.

(4) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−81.

(5) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−7.

(6) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251−4.

(7) Malmstrom, J.; Beck, M.; Schmidt, A.; Lange, V.; Deutsch, E. W.; Aebersold, R. Proteome-wide cellular protein concentrations of the

human pathogen Leptospira interrogans. *Nature* **2009**, *460* (7256), 762–5.

(8) Ludwig, C.; Aebersold, R. Getting Absolute: Determining Absolute Protein Quantitites via Selected Reaction Monitoring Mass Spectrometry. In *Quantitative Proteomics*; Eyers, C. E., Gaskell, S. J., Eds.; Royal Society of Chemistry: Cambridge, 2014; Vol. *1*, pp 80–109.

(9) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **2014**, *11* (3), 319–24.

(10) Picotti, P.; Clement-Ziza, M.; Lam, H.; Campbell, D. S.; Schmidt, A.; Deutsch, E. W.; Rost, H.; Sun, Z.; Rinner, O.; Reiter, L.; Shen, Q.; Michaelson, J. J.; Frei, A.; Alberti, S.; Kusebauch, U.; Wollscheid, B.; Moritz, R. L.; Beyer, A.; Aebersold, R. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **2013**, *494* (7436), 266–70.

(11) Li, Y. F.; Arnold, R. J.; Tang, H.; Radivojac, P. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.* **2010**, *9* (12), 6288–97.

(12) Brownridge, P.; Beynon, R. J. The importance of the digest: proteolysis and absolute quantification in proteomics. *Methods* **2011**, *54* (4), 351–60.

(13) Leon, I. R.; Schwammle, V.; Jensen, O. N.; Sprenger, R. R. Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol. Cell. Proteomics* **2013**, *12* (10), 2992–3005.

(14) Glatter, T.; Ludwig, C.; Ahrne, E.; Aebersold, R.; Heck, A. J.; Schmidt, A. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* **2012**, *11* (11), 5145–56.

(15) Gerster, S.; Kwon, T.; Ludwig, C.; Matondo, M.; Vogel, C.; Marcotte, E. M.; Aebersold, R.; Buhlmann, P. Statistical approach to protein quantification. *Mol. Cell. Proteomics* **2014**, *13* (2), 666–77.

(16) Wisniewski, J. R.; Hein, M. Y.; Cox, J.; Mann, M. A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards. *Mol. Cell. Proteomics* **2014**, *13* (12), 3497–506.

(17) Bythell, B. J.; Csonka, I. P.; Suhai, S.; Barofsky, D. F.; Paizs, B. Gas-phase structure and fragmentation pathways of singly protonated peptides with N-terminal arginine. *J. Phys. Chem. B* **2010**, *114* (46), 15092–105.

(18) Bleiholder, C.; Osburn, S.; Williams, T. D.; Suhai, S.; Van Stipdonk, M.; Harrison, A. G.; Paizs, B. Sequence-scrambling fragmentation pathways of protonated peptides. *J. Am. Chem. Soc.* **2008**, *130* (52), 17774–89.

(19) Jia, C.; Qi, W.; He, Z.; Qiao, B. Multi-stage collisionally-activated decomposition in an ion trap for identification of sequences, structures and bn -> bn-1 fragmentation pathways of protonated cyclic peptides. *Eur. Mass Spectrom.* **2006**, *12* (4), 235–45.

(20) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–48.

(21) Zhurov, K. O.; Fornelli, L.; Wodrich, M. D.; Laskay, U. A.; Tsybin, Y. O. Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis. *Chem. Soc. Rev.* **2013**, *42* (12), 5014–5030.

(22) Enke, C. G. A predictive model for matrix and analyte effects in electrospray ionization of singly-charged ionic analytes. *Anal. Chem.* **1997**, *69* (23), 4885–93.

(23) Du, L.; White, R. L. Improved partition equilibrium model for predicting analyte response in electrospray ionization mass spectrometry. *J. Mass Spectrom.* **2009**, *44* (2), 222–9.

(24) Remane, D.; Meyer, M. R.; Wissenbach, D. K.; Maurer, H. H. Ion suppression and enhancement effects of co-eluting analytes in multi-analyte approaches: systematic investigation using ultra-high-performance liquid chromatography/mass spectrometry with atmospheric-pressure chemical ionization or electrospray ionization. *Rapid Commun. Mass Spectrom.* **2010**, *24* (21), 3103–8.

(25) Cech, N. B.; Enke, C. G. Relating electrospray ionization response to nonpolar character of small peptides. *Anal. Chem.* **2000**, *72* (13), 2717–23.

(26) Nilsson, L. B.; Skansen, P. Investigation of absolute and relative response for three different liquid chromatography/tandem mass spectrometry systems; the impact of ionization and detection saturation. *Rapid Commun. Mass Spectrom.* **2012**, *26* (12), 1399–406.

(27) Osaka, I.; Takayama, M. Influence of hydrophobicity on positive- and negative-ion yields of peptides in electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **2014**, *28* (20), 2222–6.

(28) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteomics* **2011**, *10* (11), M110 003384.

(29) Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.* **2009**, *27* (2), 190–8.

(30) Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **2007**, *25* (1), 117–24.

(31) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125–31.

(32) Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **2006**, *22* (14), e481–8.

(33) Holman, S. W.; Sims, P. F. G.; Eyers, C. E. The use of selected reaction monitoring in quantitative proteomics. *Bioanalysis* **2012**, *4* (14), 1763–1786.

(34) Beynon, R. J.; Doherty, M. K.; Pratt, J. M.; Gaskell, S. J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2005**, *2* (8), 587–9.

(35) Brownridge, P.; Holman, S. W.; Gaskell, S. J.; Grant, C. M.; Harman, V. M.; Hubbard, S. J.; Lanthaler, K.; Lawless, C.; O'Cualain, R.; Sims, P.; Watkins, R.; Beynon, R. J. Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy. *Proteomics* **2011**, *11* (15), 2957–70.

(36) Lawless, C.; Holman, S. W.; Brownridge, P.; Lanthaler, K.; Harman, V. M.; Watkins, R.; Hammond, D. E.; Miller, R. L.; Sims, P. F.; Grant, C. M.; Eyers, C. E.; Beynon, R. J.; Hubbard, S. J. Direct and Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. *Mol. Cell. Proteomics* **2016**, *15* (4), 1309–22.

(37) Muntel, J.; Boswell, S. A.; Tang, S.; Ahmed, S.; Wapinski, I.; Foley, G.; Steen, H.; Springer, M. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). *Mol. Cell. Proteomics* **2015**, *14* (2), 430–40.

(38) Brownridge, P.; Lawless, C.; Payapilly, A. B.; Lanthaler, K.; Holman, S. W.; Harman, V. M.; Grant, C. M.; Beynon, R. J.; Hubbard, S. J. Quantitative analysis of chaperone network throughput in budding yeast. *Proteomics* **2013**, *13* (8), 1276–91.

(39) Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **2014**, *13* (12), 3698–708.

(40) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–72.

(41) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10* (4), 1794–805.

(42) Lawless, C.; Hubbard, S. J. Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *OMICS* **2012**, *16* (9), 449−56.

(43) Krokhin, O. V.; Spicer, V. Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem.* **2009**, *81* (22), 9522−30.

(44) Andrews, P. C.; Arnott, D. P.; Gawinowicz, M. A.; Kowalak, J. A.; Lane, W. S.; Lilley, K. S.; Martin, L. T.; Stein, S. ABRF-sPRG 2006 study: a proteomics standard. ABRF 2006: Long Beach, CA, 2006.

(45) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009**, *138* (4), 795−806.

(46) Kawashima, S.; Ogata, H.; Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **1999**, *27* (1), 368−9.

(47) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R., 3rd Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (5), 1243−8.

(48) Searle, B. C.; Egertson, J. D.; Bollinger, J. G.; Stergachis, A. B.; MacCoss, M. J. Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. *Mol. Cell. Proteomics* **2015**, *14* (9), 2331−40.

(49) Boyd, R.; Somogyi, A. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (8), 1275−8.

(50) Kapp, E. A.; Schutz, F.; Reid, G. E.; Eddes, J. S.; Moritz, R. L.; O'Hair, R. A.; Speed, T. P.; Simpson, R. J. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **2003**, *75* (22), 6251−64.

(51) Lanucara, F.; Lee, D. C.; Eyers, C. E. Unblocking the sink: improved CID-based analysis of phosphorylated peptides by enzymatic removal of the basic C-terminal residue. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (2), 214−25.

(52) Gucinski, A. C.; Dodds, E. D.; Li, W.; Wysocki, V. H. Understanding and exploiting Peptide fragment ion intensities using experimental and informatic approaches. *Methods Mol. Biol.* **2010**, *604*, 73−94.

(53) Tang, K.; Smith, R. D. Physical/chemical separations in the break-up of highly charged droplets from electrosprays. *J. Am. Soc. Mass Spectrom.* **2001**, *12* (3), 343−7.

(54) Cowan, R.; Whittaker, R. G. Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Pept. Res.* **1990**, *3* (2), 75−80.

(55) Cech, N. B.; Krone, J. R.; Enke, C. G. Predicting electrospray response from chromatographic retention time. *Anal. Chem.* **2001**, *73* (2), 208−13.

(56) Abaye, D. A.; Pullen, F. S.; Nielsen, B. V. Peptide polarity and the position of arginine as sources of selectivity during positive electrospray ionisation mass spectrometry. *Rapid Commun. Mass Spectrom.* **2011**, *25* (23), 3597−608.

(57) de Godoy, L. M.; Olsen, J. V.; de Souza, G. A.; Li, G.; Mortensen, P.; Mann, M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **2006**, *7* (6), R50.

(58) Michalski, A.; Cox, J.; Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785−93.

(59) Schliekelman, P.; Liu, S. Quantifying the effect of competition for detection between coeluting peptides on detection probabilities in mass-spectrometry-based proteomics. *J. Proteome Res.* **2014**, *13* (2), 348−61.

(60) Ghaemmaghami, S.; Huh, W. K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. Global analysis of protein expression in yeast. *Nature* **2003**, *425* (6959), 737−41.

(61) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002**, *1* (11), 845−67.

(62) Weiss, M.; Schrimpf, S.; Hengartner, M. O.; Lercher, M. J.; von Mering, C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **2010**, *10* (6), 1297−306.

DOI: 10.1021/acs.jproteome.6b00048