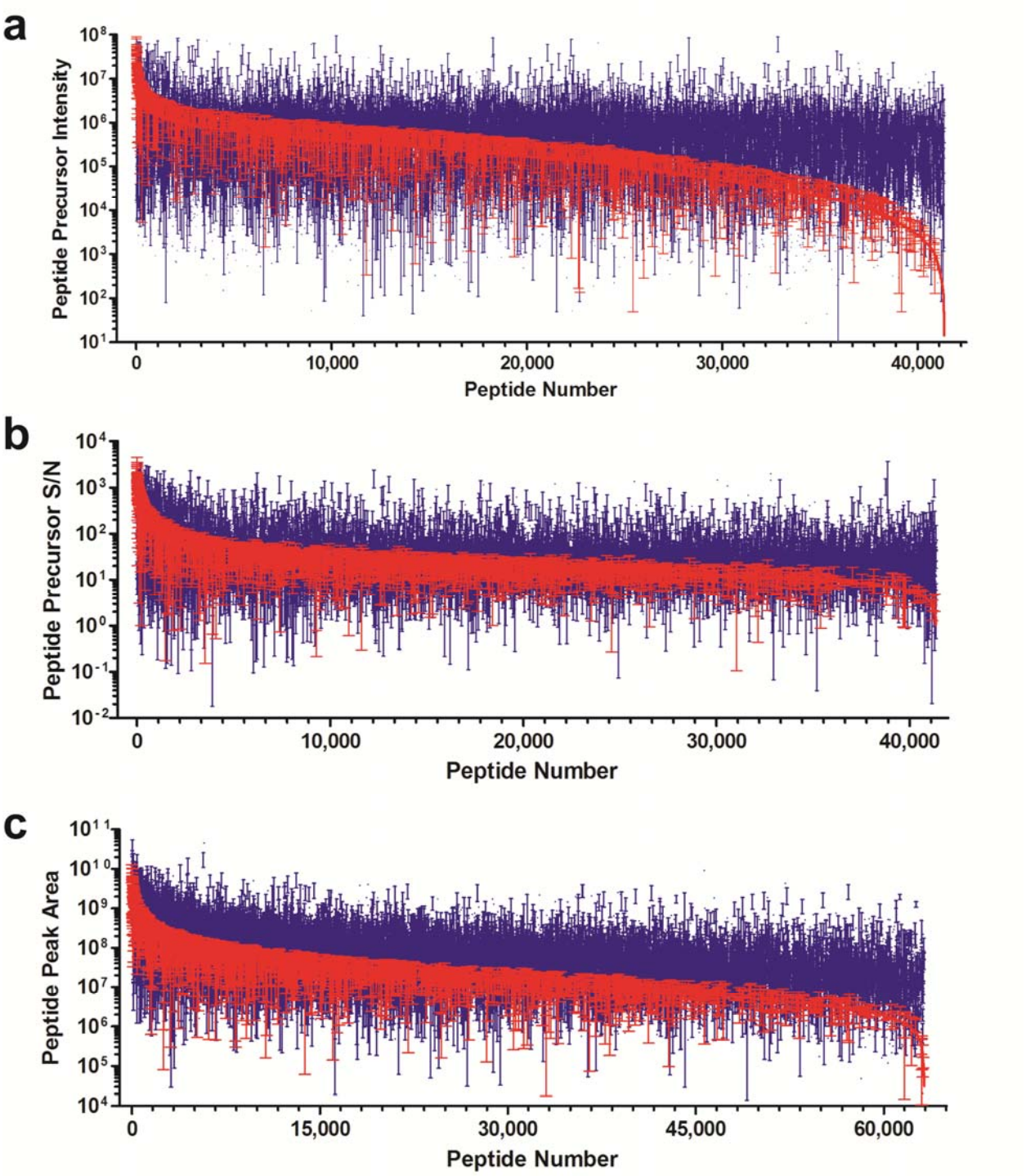
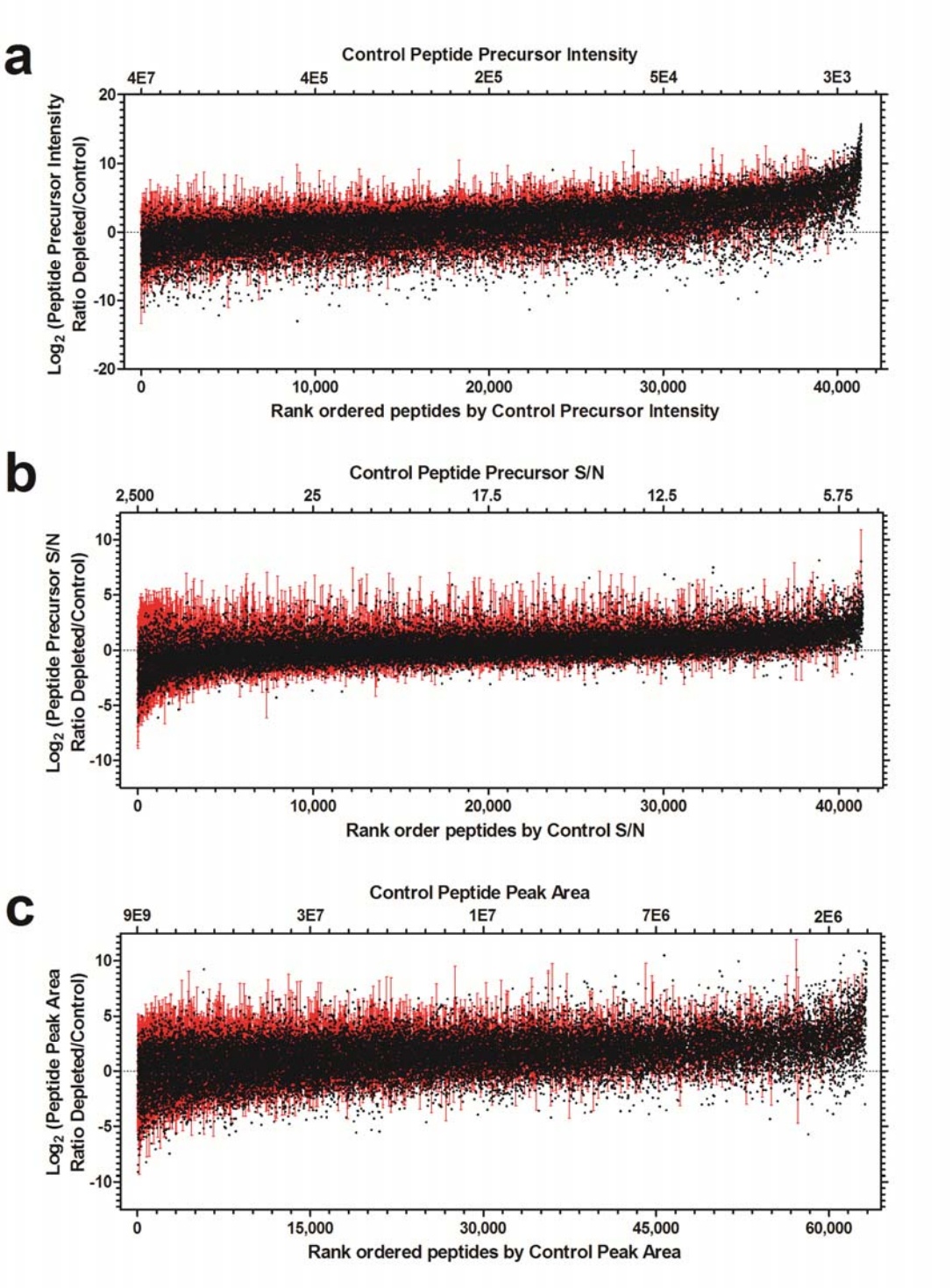
Supplementary Figures and Notes for *Addendum: Digestion and depletion of abundant proteins improves proteomic coverage*

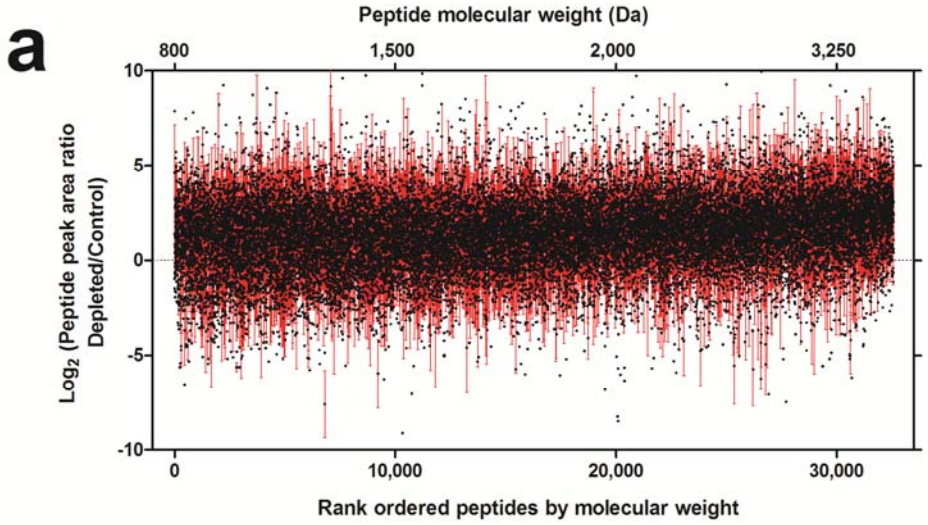
Bryan R. Fonslow, Mark S. Hixon, Benjamin D. Stein, Kristofor J. Webb, Tao Xu, Jeong Choi, Sung Kyu Park, and John R. Yates III,*

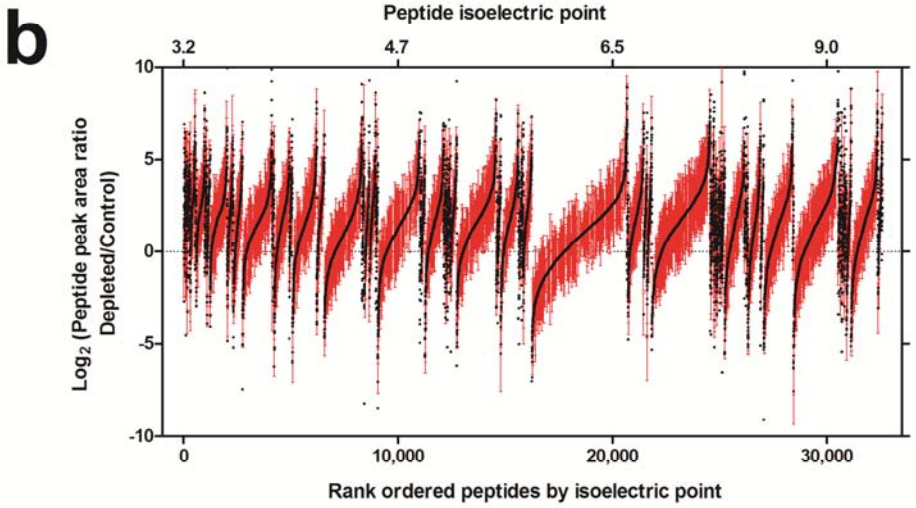


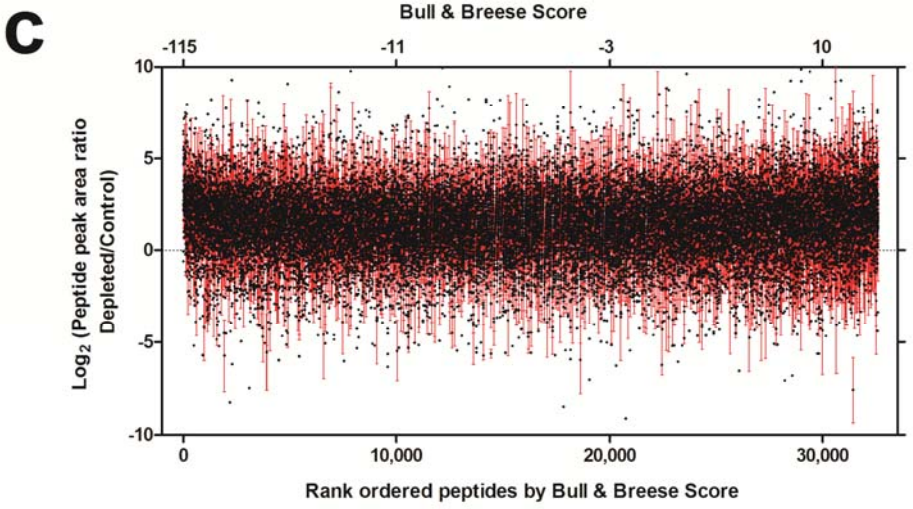
Supplementary Figure 1: Rank abundance plots of peptide (a) precursor intensities, (b) precursor S/N ratios, and (c) chromatographic peak areas from triplicate control (red) and DigDeAPr (blue) runs. Standard deviations are represented by error bars. Precursor intensities, S/N ratios, and peak areas were considered for the highest scoring peptides within a MudPIT run.

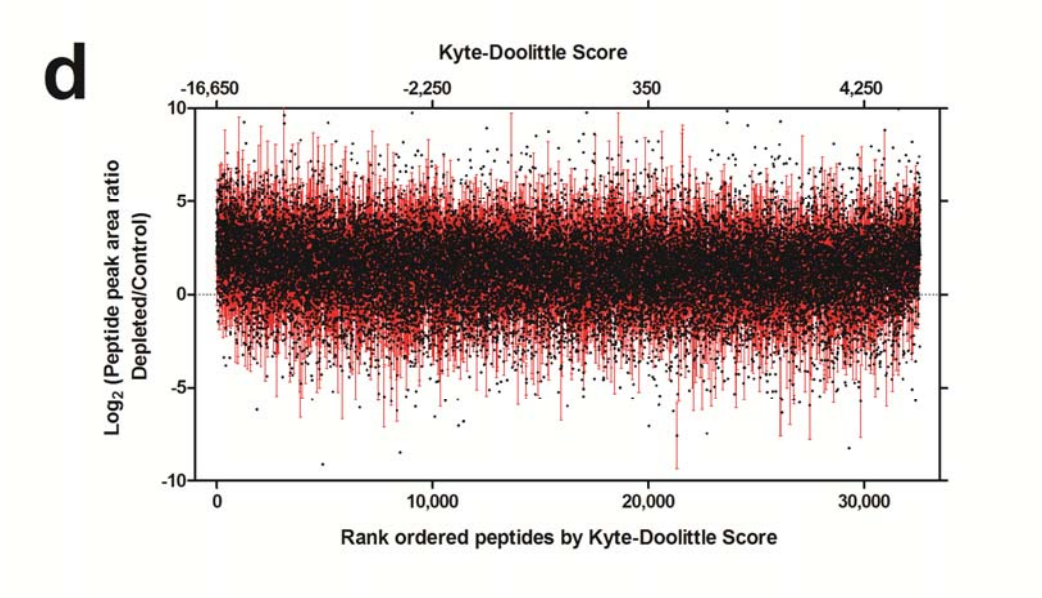


Supplementary Figure 2: Log_2 peptide ratio (black) rank abundance plots of peptide (a) precursor intensities, (b) precursor S/N ratios, and (c) chromatographic peak areas from triplicate control and DigDeAPr runs. Standard deviations are represented by error bars (red). Precursor intensities, S/N ratios, and peak areas were considered for the highest scoring peptide match for the same sequence within a MudPIT run.

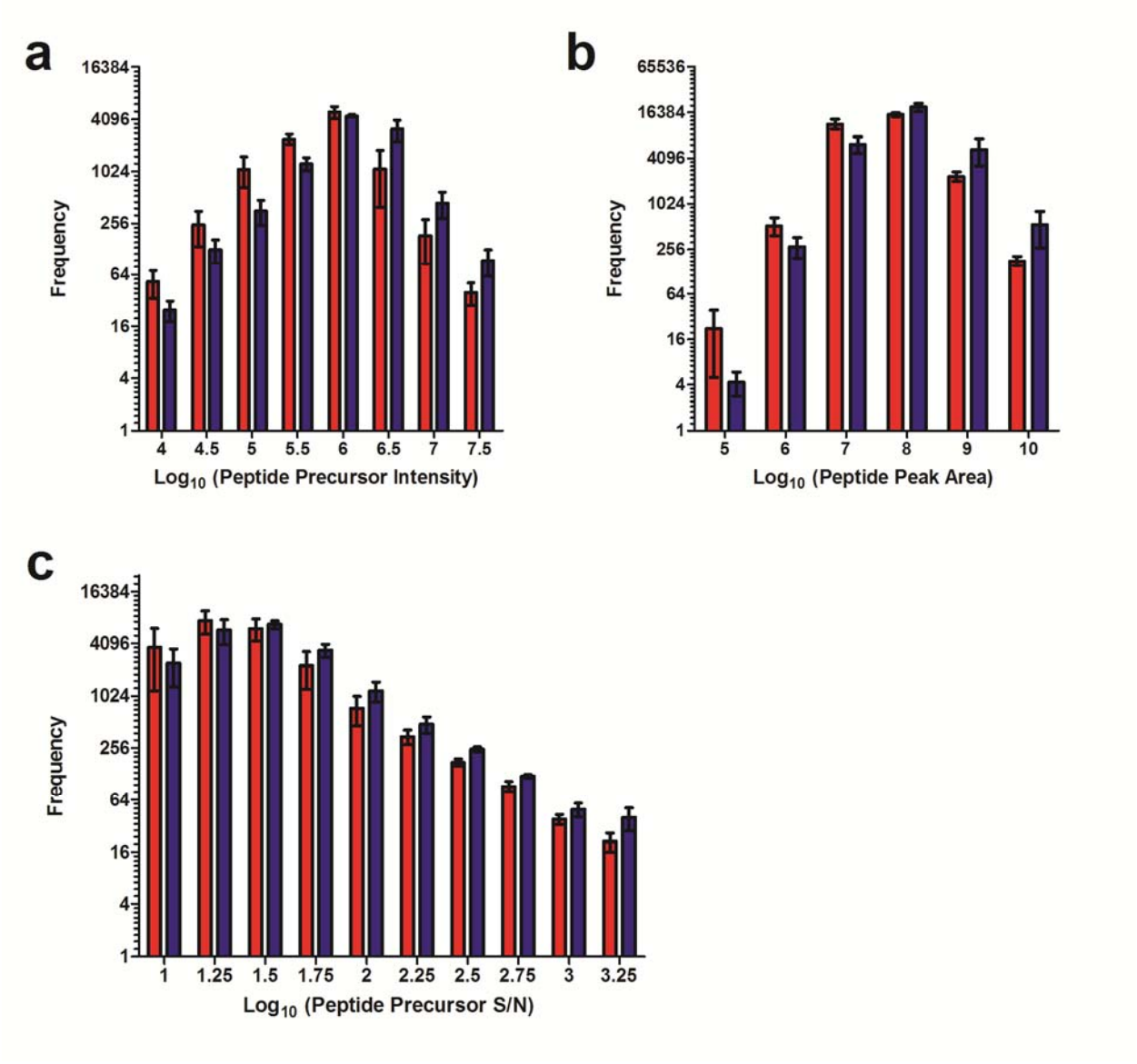

a
Peptide molecular weight (Da)
800
1,500
2,000
3,250
10
5
0
-5
-10
Log2 (Peptide peak area ratio Depleted/Control)
0
10,000
20,000
30,000
Rank ordered peptides by molecular weight


b
Peptide isoelectric point
3.2
4.7
6.5
9.0
10
5
0
-5
-10
Log2 (Peptide peak area ratio Depleted/Control)
0
10,000
20,000
30,000
Rank ordered peptides by isoelectric point

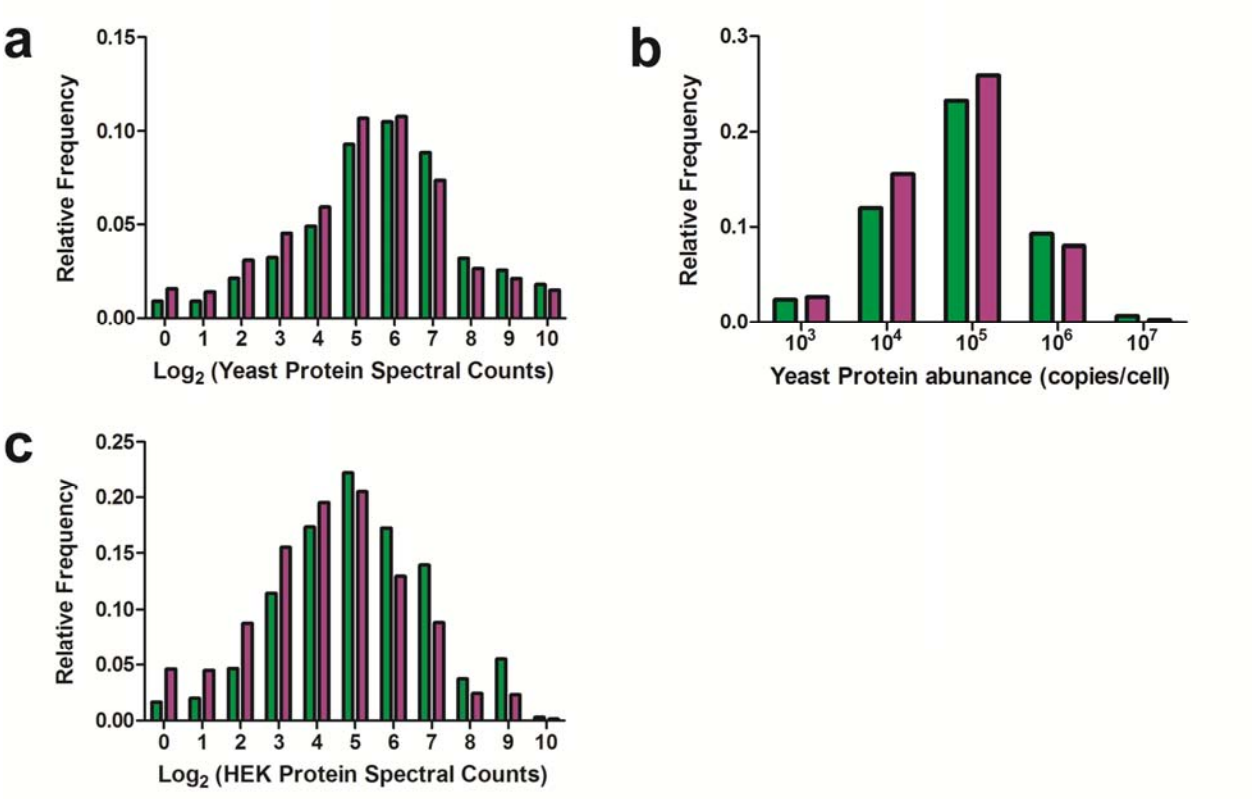

c
Bull & Breese Score
-115
-11
-3
10
10
5
0
-5
-10
Log2 (Peptide peak area ratio Depleted/Control)
0
10,000
20,000
30,000
Rank ordered peptides by Bull & Breese Score



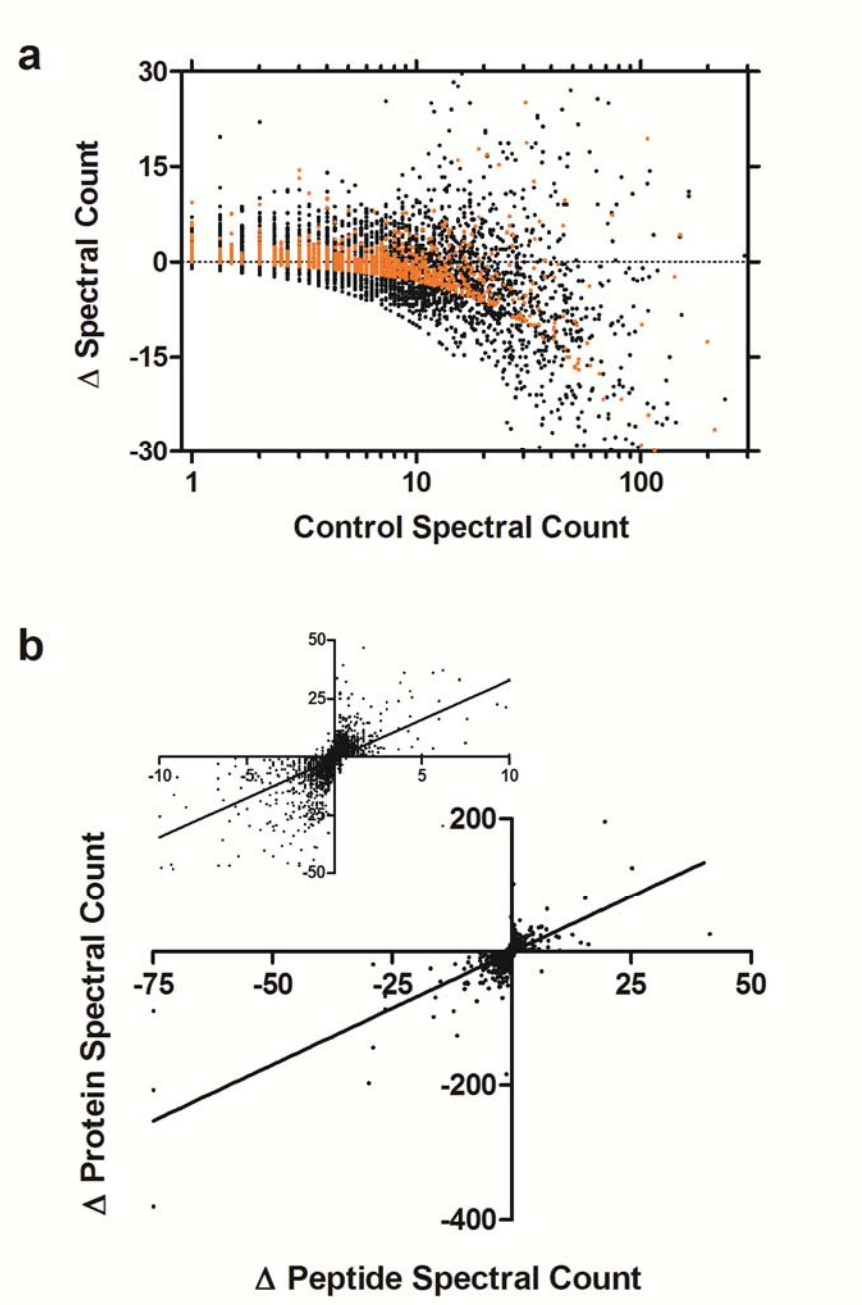
Supplementary Figure 3: Log_2 peptide peak area ratio (black) rank abundance plots of peptide (a) molecular weight, (b) isoelectric point, (c) Bull and Breese score, and (d) Kyte-Doolittle Score from triplicate control and DigDeAPr runs. Standard deviations are represented by error bars (red). Peak areas were considered for the highest scoring peptide match with the same sequence within a MudPIT run.



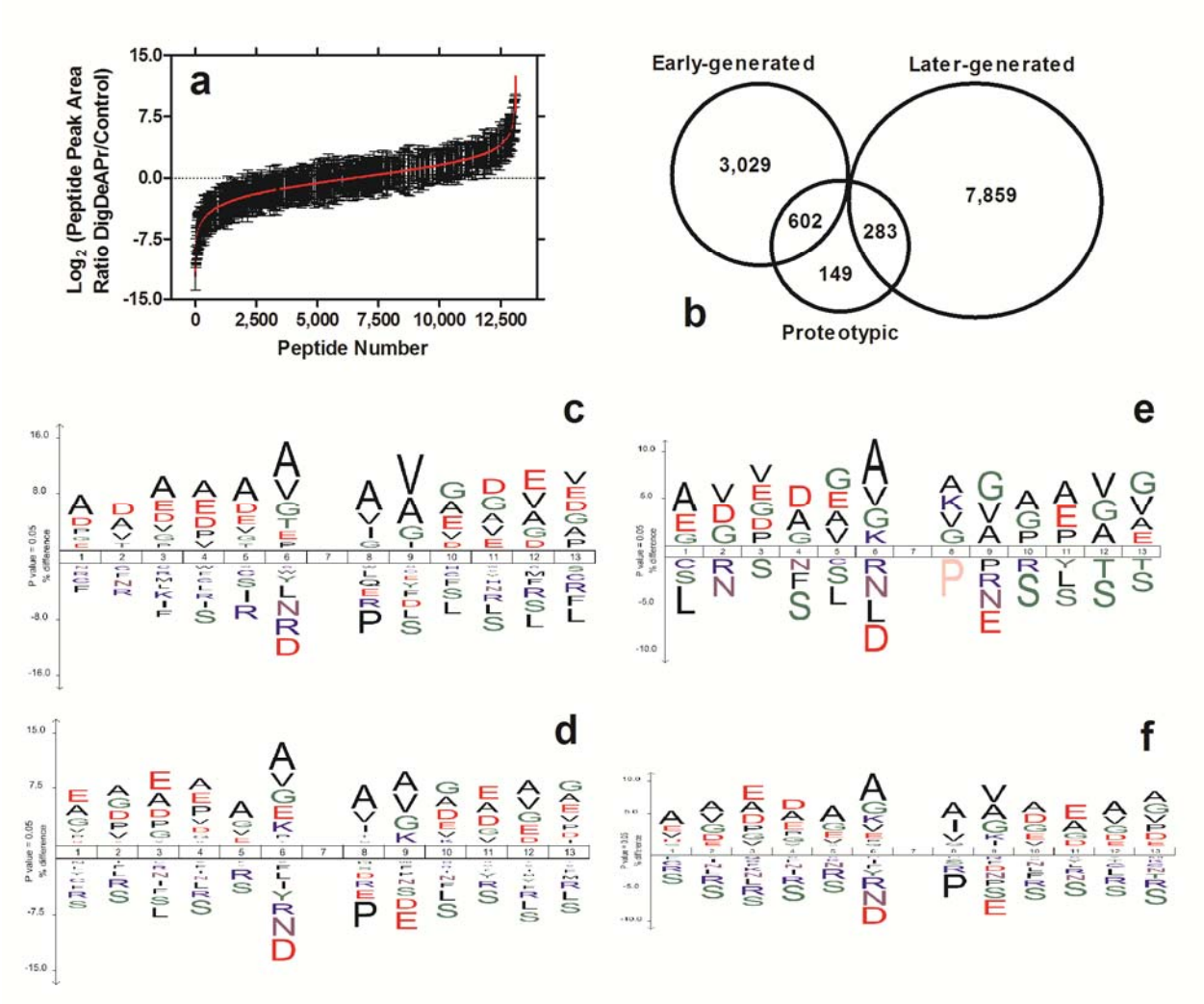
Supplementary Figure 4: HEK cell lysate peptide (a) precursor intensity, (b) S/N, and (c) peak area histograms comparing triplicate control (red) and DigDeAPr (blue) runs with error bars representing standard deviation. A systematic increase in peptide precursor intensity, S/N, and peak area were found for all peptides identified in DigDeAPr runs relative to control runs.



Supplementary Figure 5: Distributions of early-generated peptides (green, $\log_2$ peptide peak area $\leq$ -1) and later-generated peptides (purple, $\log_2$ peptide peak area $\geq$ 1) based on (a) yeast protein spectral counts, (b) yeast absolute protein abundances from Ghaemmaghami *et al*,[1] and (c) HEK protein spectral counts.



Supplementary Figure 6: Correlation of HEK peptide to HEK protein spectral count changes. (a) Correlation plot of protein (black) and peptide (orange) spectral counts changes from DigDeAPr to control protein spectral counts. Δ spectral counts are DigDeAPr minus control runs. (b) Correlation plot of highest single peptide spectral count changes to the total protein spectral count changes. The greatest peptide spectral count changes, positive or negative, were used with the same directional change as the protein spectral count. The fit line can be described by $y = 3.36x - 0.78$; $R^2 = 0.57$.



Supplementary Figure 7: (a) Distribution of quantified yeast peptide ratios using label-free peptide peak area measurements. (b) Venn diagram of early-generated and later-generated yeast peptides from analysis herein and proteotypic peptides from Mallick *et al*.[2] Tryptic site motif analysis of yeast (c) proteotypic peptides (n = 1,627), (d) early-generated and depleted peptides based on chromatographic ratios (n = 5,733), (e) peptides depleted on average by 5 or more spectral counts (n = 774), and (f) peptides depleted on average by 1 or more spectral counts (n =2,299) versus Uniprot *S. cerevisiae* database comparing regional motifs.

Supplementary Note 1: Theory of Digestion and Depletion. Many recent studies have begun to re-illustrate the importance, benefits, and challenges of protease digestions on protein identification and quantitation. The use of multiple proteases in different combinations have been shown to improve proteomic coverage[3-8] and digestion efficiency.[9] Different proteases generate different populations of peptides to improve proteome coverage. While this improves sequence coverage of proteins, sequence biases among proteins digested with different proteases have been shown to affect protein quantitation with spectral counting.[10] Additionally, not all sites for a single protease are equal. For instance, trypsin cleaves C-terminal to Lys and Arg, but when followed by Pro, has proximal positive charges such as successive Lys/Arg, or has several proximal Glu/Asp residues, these sites are less likely to be cleaved, as described by the Keil rules.[11] Large data sets of peptides generated by shotgun proteomics experiments have facilitated more comprehensive analyses and prediction of missed cleavage events. An information theory analysis of missed cleavages found a modest trypsin inhibition by proximal Met, Ser, and Gly to Lys/Arg in addition to known rules.[12] A machine learning-based decision tree approach on a more comprehensive data set was used to assign probabilities to cleavage events.[13] These missed cleavage analysis generally rely on peptide data sets that were digested to completion, lacking a consideration of the dynamics and kinetics of the proteolytic events. Time-course analyses begin to illustrate the differential cleavage rates of proteolytic sites and the differential generation of peptides. A gel-based time-course analysis of a single protein, human albumin, showed early- and later-generated peptides.[14] A time-course analysis with peptides and mass spectrometry facilitated calculation of K_M and V_{max} for trypsin.[15]

A similar concept was applied globally to the analysis of human apoptotic caspase kinetic efficiencies (k_{cat}/K_M) by mass spectrometry.[16] Surprisingly, very little attention has been given to the kinetics of proteolytic digestion to whole cell lysates. An understanding and application of these concepts to proteolytic digestions of whole cell lysates should benefit shotgun proteomic analyses. In this case, recognizing fast and slow tryptic sites and how they contribute to the generation and separation of early- and later-generated peptides is pertinent.

In order to describe how digestion and depletion of early-generated peptides affect our proteomic analysis, we provide an approximate solution which describes the mole fractions of multiple competing substrates based on time (t) and site selectivity (k_{cat}/K_M). These parameters are illustrated by the well-known partition equation for two competing substrates:[17]

$$\frac{v_A}{v_b} = \frac{\left(\frac{k_{cat}}{K_M}\right)_A [A]}{\left(\frac{k_{cat}}{K_M}\right)_B [B]} \qquad (1)$$

From Fersht, “The important conclusion is that the specificity, in the sense of discrimination between two competing substrates, is determined by the ratios of k_{cat}/K_M and not by K_M alone.” This fact was overlooked in our initial attempt to determine a mechanism, as we considered only K_M and protein concentration.[18] With this new insight, we have expanded equation 1 to a more general solution describing three or more competing substrates to better understand this phenomenon during digestion and depletion where thousands of substrates are competing.

Defining the total enzyme ($[E_T]$) as the sum of free concentration ($[E_{Free}]$) and substrate complexes concentrations ($[E_A]$, $[E_B]$, $[E_C]$, and $[E_n]$) of substrates A, B, C, and n (representing more than three substrates), the total enzyme can be expressed as a function of steady state concentrations:

$$[E_T] = \left(1 + \frac{[A]}{K_A} + \frac{[B]}{K_B} + \frac{[C]}{K_C} + \frac{[n]}{K_n}\right)[E] \tag{2}$$

And the free enzyme as:

$$[E] = \frac{[E_T]}{\left(1 + \frac{[A]}{K_A} + \frac{[B]}{K_B} + \frac{[C]}{K_C} + \frac{n}{K_n}\right)} \tag{3}$$

Inserting the free enzyme concentration equation 3 into the rate equation for substrate A:

$$\frac{d[A]}{dt} = -\frac{k_{cat}^{A}[A]}{K_A}[E] \tag{4}$$

yields an equation in terms of total enzyme (E_T) which can be quantified, unlike the free enzyme concentration:

$$\frac{d[A]}{dt} = \frac{-\frac{k_{cat}^{A}[A]}{K_A}}{\left(1 + \frac{[A]}{K_A} + \frac{[B]}{K_B} + \frac{[C]}{K_C} + \frac{[n]}{K_n}\right)}[E_T] \tag{5}$$

The same procedure can be repeated to create an equation that represents the total substrate $[S_T]$ consumption:

$$\frac{d[S_T]}{dt} = \frac{-\frac{k_{cat}^A[A]}{K_A} - \frac{k_{cat}^B[B]}{K_B} - \frac{k_{cat}^C[C]}{K_C} - \frac{k_{cat}^n[n]}{K_n}}{\left(1+\frac{[A]}{K_A}+\frac{[B]}{K_B}+\frac{[C]}{K_C}+\frac{[n]}{K_n}\right)}[E_T] \quad (6)$$

Comparing the rate of consumption of substrate A to the total substrate consumption:

$$\frac{\frac{d[A]}{dt}}{\frac{d[S_T]}{dt}} = \frac{\frac{-\frac{k_{cat}^A[A]}{K_A}}{\left(1+\frac{[A]}{K_A}+\frac{[B]}{K_B}+\frac{[C]}{K_C}+\frac{[n]}{K_n}\right)}[E_T]}{\frac{-\frac{k_{cat}^A[A]}{K_A} - \frac{k_{cat}^B[B]}{K_B} - \frac{k_{cat}^C[C]}{K_C} - \frac{k_{cat}^n[n]}{K_n}}{\left(1+\frac{[A]}{K_A}+\frac{[B]}{K_B}+\frac{[C]}{K_C}+\frac{[n]}{K_n}\right)}[E_T]} \quad (7)$$

Multiplying through by E_T and canceling both denominator terms:

$$\frac{\frac{d[A]}{dt}}{\frac{d[S_T]}{dt}} = \frac{-\frac{k_{cat}^A}{K_A}[E_T][A]}{\left(-\frac{k_{cat}^A[E_T]}{K_A}[A]-\frac{k_{cat}^B[E_T]}{K_B}[B]-\frac{k_{cat}^C[E_T]}{K_C}[C]-\frac{k_{cat}^n[E_T]}{K_n}[n]\right)} \quad (8)$$

And writing in enzymology shorthand, $k_{cat} \times [E_T]$ is expressed as (V), and the Michaelis constant is expressed as (K) for a specificity constant (V/K) yields:

$$\frac{v_A}{v_T} = \frac{\left(\frac{V}{K}\right)_A[A]}{\left(\frac{V}{K}\right)_A[A]+\left(\frac{V}{K}\right)_B[B]+\left(\frac{V}{K}\right)_C[C]+\left(\frac{V}{K}\right)_n[n]} \quad (9)$$

This equation begins to illustrate how the rate of consumption of substrate A is dependent on the rate of consumption of competing substrates. The consumption of

substrates A, B, C, or n, whichever occurs first, is then depleted during the spin-filter step in DigDeAPr. Thus, this is similar to our initial mechanistic proposal where we described abundant proteins as inhibitors for digestion of low abundance proteins. However, from this derivation we can now conclude that fast tryptic cleavage sites compete with slow tryptic cleavage sites based on the abundance of a protein and on their relative specificity (V/K) of particular cleavage site (e.g. $[A]$).

To derive a more quantitative description of substrate consumption and relative depletion at any particular assay time we must integrate the rate equation. For clarity it is best to express competing substrates in terms of the mole fraction (χ) of a particular substrate relative to the total substrate pool:

$$\frac{\frac{d[A]}{dt}}{\frac{d[S_T]}{dt}} = \frac{\left(\frac{V}{K}\right)_A [A]}{\left(\frac{V}{K}\right)_A [A] + \left(\frac{V}{K}\right)_B [B] + \left(\frac{V}{K}\right)_C [C] + \left(\frac{V}{K}\right)_n [n]} \tag{10}$$

Expressing the individual substrates as mole fractions of the total gives:

$$\frac{\frac{d\chi_A[S_T]}{dt}}{\frac{d[S_T]}{dt}} = \frac{\left(\frac{V}{K}\right)_A \chi_A[S_T]}{\left(\frac{V}{K}\right)_A \chi_A[S_T] + \left(\frac{V}{K}\right)_B \chi_B[S_T] + \left(\frac{V}{K}\right)_C \chi_C[S_T] + \left(\frac{V}{K}\right)_n \chi_n[S_T]} \tag{11}$$

The differential equation is solved by separation of variables and then conducting integration by parts for the denominator:

$$\frac{-\frac{1}{\left(\frac{V}{K}\right)_A}\int\frac{1}{\chi_A}d\chi_A[S_T]=\int dt}{-\frac{1}{\left(\frac{V}{K}\right)_A}\int\frac{1}{\chi_A}d\chi_A[S_T]=\int dt-\frac{1}{\left(\frac{V}{K}\right)_B}\int\frac{1}{\chi_B}d\chi_B[S_T]=\int dt-\frac{1}{\left(\frac{V}{K}\right)_C}\int\frac{1}{\chi_C}d\chi_C[S_T]=\int dt-\frac{1}{\left(\frac{V}{K}\right)_n}\int\frac{1}{\chi_n}d\chi_n[S_T]=\int dt} \tag{12}$$

Obtaining the definite integral:

$$\frac{-\frac{1}{\left(\frac{V}{K}\right)_A}\ln\frac{\chi_A[S_T]}{(\chi_A[S_T])_{t=0}}=t}{-\frac{1}{\left(\frac{V}{K}\right)_A}\ln\frac{\chi_A[S_T]}{(\chi_A[S_T])_{t=0}}=t-\frac{1}{\left(\frac{V}{K}\right)_B}\ln\frac{\chi_B[S_T]}{(\chi_B[S_T])_{t=0}}=t-\frac{1}{\left(\frac{V}{K}\right)_C}\ln\frac{\chi_C[S_T]}{(\chi_C[S_T])_{t=0}}=t-\frac{1}{\left(\frac{V}{K}\right)_n}\ln\frac{\chi_n[S_T]}{(\chi_n[S_T])_{t=0}}=t} \tag{13}$$

and taking the antilog yields a quantitative expression for the mole fraction of substrate A (χ_A) at any time in terms of the mole fraction of other competing substrates (χ_B, χ_C, and χ_n) at the start of the digestion (t = 0):

$$\chi_A=\frac{(\chi_A)_{t=0}e^{\left(-\frac{V}{K}\right)_A t}}{(\chi_A)_{t=0}e^{\left(-\frac{V}{K}\right)_A t}+(\chi_B)_{t=0}e^{\left(-\frac{V}{K}\right)_B t}+(\chi_C)_{t=0}e^{\left(-\frac{V}{K}\right)_C t}+(\chi_n)_{t=0}e^{\left(-\frac{V}{K}\right)_n t}} \tag{14}$$

Like rate equation (9) where the velocity of substrate A consumption is expressed as a relation to the total substrate velocity (v_A/v_T), equation (14) expresses the mole fraction of substrate A in terms of the mole fractions of all other competing substrates. Inclusion of χ_n, extends the equation to be descriptive of a complex proteome mixture with n different tryptic motifs:

$$\chi_A=\frac{(\chi_A)_{t=0}e^{\left(-\frac{V}{K}\right)_A t}}{(\chi_A)_{t=0}e^{\left(-\frac{V}{K}\right)_A t}+(\chi_B)_{t=0}e^{\left(-\frac{V}{K}\right)_B t}+(\chi_C)_{t=0}e^{\left(-\frac{V}{K}\right)_C t}+(\chi_n)_{t=0}e^{\left(-\frac{V}{K}\right)_n t}} \tag{15}$$

From equation 15 we see that at time zero we have the natural abundance of substrates (χ_A). As time (t) progresses, generation of a particular substrate depends solely on the specificity constant of that substrate, in the numerator, relative to all competing substrate specificity constants, in the denominator. With this in mind, equation 15 illustrates that defined populations of peptides will be generated over time based on their specificity constants. In order to understand the final abundance of a substrate, as a mole fraction, after limited digestion and depletion ($\chi_{A,depleted}$), we must consider removal of substrates ($\chi_{A,depletion}$) at the depletion time (t_d) in the context of the substrates present ($\chi_{A,complete}$) after the complete digestion time (t_c):

$$\chi_{A,depletion} = \frac{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_d}}{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_d} + (\chi_B)_{t=0} e^{\left(-\frac{V}{K}\right)_B t_d} + (\chi_C)_{t=0} e^{\left(-\frac{V}{K}\right)_C t_d} + (\chi_n)_{t=0} e^{\left(-\frac{V}{K}\right)_n t_d}} \quad (16)$$

$$\chi_{A,complete} = \frac{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_c}}{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_c} + (\chi_B)_{t=0} e^{\left(-\frac{V}{K}\right)_B t_c} + (\chi_C)_{t=0} e^{\left(-\frac{V}{K}\right)_C t_c} + (\chi_n)_{t=0} e^{\left(-\frac{V}{K}\right)_n t_c}} \quad (17)$$

The depleted mole fraction ($\chi_{A,depleted}$) is then the difference between the final, complete digestion mole fraction ($\chi_{A,complete}$) and the mole fraction at the depletion step ($\chi_{A,depletion}$):

$$\chi_{A,depleted} = \chi_{A,complete} - \chi_{A,depletion} \quad (18)$$

Substituting equations 16 and 17 into equation 18 yields:

$$\chi_{A,depleted} = \frac{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_c}}{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_c} + (\chi_B)_{t=0} e^{\left(-\frac{V}{K}\right)_B t_c} + (\chi_C)_{t=0} e^{\left(-\frac{V}{K}\right)_C t_c} + (\chi_n)_{t=0} e^{\left(-\frac{V}{K}\right)_n t_c}} - \frac{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_d}}{(\chi_A)_{t=0} e^{\left(-\frac{V}{K}\right)_A t_d} + (\chi_B)_{t=0} e^{\left(-\frac{V}{K}\right)_B t_d} + (\chi_C)_{t=0} e^{\left(-\frac{V}{K}\right)_C t_d} + (\chi_n)_{t=0} e^{\left(-\frac{V}{K}\right)_n t_d}} \quad (19)$$

Equation 19 is an approximate solution describing protein depletion. For simplicity 19 assumes one cleavage site per protein; alternatively one could say it presents a composite (V/K) of all *exposed* cleavage sites on a particular protein. If one wishes to explicitly account for multiple exposed cleavage sites, each (V/K) term in 19 becomes the sum of the individual protein's cleavage sites specificity constants Σ(V/K). Lastly, 19 is an approximate solution since it does not account for the processive nature of protein cleavage i.e., one cleavage event will expose interior cleavage sites to proteolysis. In effect this causes the denominator to grow with time as new substrate cleavage sites are being created, adding to the competition for a specific cleavage site. As a consequence, the rate of a specific protein's cleavage is dampened. Nevertheless, 19 provides a useful description of DigDeAPr proteolysis kinetics and illustrates how specific (V/K) values determine enrichment or depletion over time.

Supplementary Note 2: Peptide abundances are equalized through depletion. Considering a peptide-centric depletion and enrichment mechanism based on cleavage site specificities (V/K), peptide abundances would be expected to change as illustrated in Figure 1a. Comparison of peptide abundance metrics from control and DigDeAPr runs facilitates empirical validation of this mechanism. We considered peptide precursor intensities, signal-to-noise ratios (S/N), and chromatographic peak areas

(Supplementary Figure 1) as relative measures of peptide abundance.[19-21] Only peptides common to both control and DigDeAPr runs were considered in these analyses for direct comparison. Rank abundance plotting of control peptide metrics establishes the “natural abundance” of peptides, as measured by the response of the mass spectrometer, denoted in Figure 1a. The “DigDeAPr abundance” of the corresponding peptides of the same sequence and charge state from DigDeAPr runs are plotted in the same order as the control peptides rank order abundance. By maintaining the same order as control peptides, the DigDeAPr abundance of peptides can be directly compared. Notably in all peptide rank abundance plots, similar trends are observed to the theoretical curves in Figure 1a. Peptides of low abundance are dramatically enriched in DigDeAPr runs, equalizing peptide metrics across all peptides. The equalization of these peptide metrics normalizes and raises the quality of peptide measurements relevant for peptide and protein quantitation. Correlation plots of these trends are shown in Supplementary Figure 2 to further illustrate and quantify these changes. Since the correlation plots are still plotted in rank order of control peptide abundance metrics, the $\log_2$ ratio correlation plots also clearly illustrate that peptides of high abundance are depleted and peptides of low abundance are enriched. Similar rank order plots of peptide physicochemical properties isoelectric point and hydrophobicity (Supplementary Figure 3b-d) do not show trends; although, a slight enrichment of larger peptides is observed (Supplementary Figure 3a), which might be expected from the 10K MWCO spin-filter step. When abundance metrics from all of the peptides identified are compared, and not just those common to control and DigDeAPr runs, more can be learned about the distributions of peptides identified. Histograms of

all peptide relative abundance metrics are shown in Supplementary Figure 4. The distribution of peptide abundance has been systematically increased by all three metrics. These results clearly illustrate that peptide abundance changes are the critical factor in the success of the DigDeAPr method.

Supplementary Note 3: Early- and later-generated peptides are separated during the depletion step. As described in the theory of digestion and depletion section, we expect the generation and depletion of peptides to be dependent on the specificity constant (V/K) of tryptic sites. In our limited digestion and depletion strategy, an early-generated peptide would be one that is smaller than 10 kDa prior to the MWCO spin-filter depletion step. Assuming an average amino acid mass of 100 Da, this essentially requires two fast tryptic sites (i.e. high V/K) to be within 100 amino acids of each other for the peptide to be depleted. Later-generated peptides are those created after the 10 kDa depletion step. Shotgun proteomics generally identifies peptides less than 30 amino acids in length or less than 3 kDa. Thus, quantitative comparison of peptide chromatographic ratios in our control and DigDeAPr runs should allow for identification of peptides that are generated early and later. Early-generated peptides that are less than 10 kDa were depleted by the 10K MWCO spin-filter step, and thus should have negative $\log_2$ ratios in comparison to control runs. Similarly, peptides that were later-generated were not depleted by the 10K MWCO spin-filter step and thus should be enriched by the use of 10 times as much starting material. These enriched peptides should have positive $\log_2$ ratios in comparison to control runs.

From data in Fonslow *et al.* with triplicate control and DigDeAPr runs, we were able to measure and compare label-free peak area ratios for 13,628 and 13,112

peptides from HEK (Figure 2a) and yeast (Supplementary Figure 7a) cells, respectively. We observed a range of peptide $\log_2$ ratio changes from -14 to 10 for HEK peptides and -13 to 12 for yeast peptides. Direct peptide ratio measurements with isotopic labeling are no doubt more accurate and reproducible, but tend to underestimate large ratio changes.[22] Thus, with replicates, calculated error, and high comprehensiveness, our label-free peptide ratio measures are appropriate for these analyses. Notably, both showed a remarkable similar range and trend. In both cases about a third of the quantified peptides (~ 4,500) could be considered early-generated ($\log_2$ ratio $\leq$ -1).

Supplementary Note 4: Early-generated peptides with fast cleavage motifs are depleted and later-generated peptides with slow cleavage motifs are enriched. The proximity of fast cleavage sites will define which peptides will be early-generated and depleted during the 10K MWCO spin-filter step. Quantitative analysis of early-generated peptide tryptic cleavage sites between control and DigDeAPr runs should further validate the depletion mechanism of early-generated peptides. Similarly, quantitative analysis of later-generated peptide tryptic cleavage sites should further validate the enrichment of later-generated peptides. We used the unchanged HEK tryptic and missed cleaved peptides as a negative background for both tryptic and missed cleavage analyses, respectively. These comparisons allow for extraction of motifs that are $\log_2$ ratio changed among the peptides that were identified, not just present within the human proteome. Using the depleted, early-generated HEK peptides we found a motif that is enriched in small, uncharged residues at P1' (Figure 2b), known to have high tryptic specificities, and depleted in charged residues known to have lower specificity.[11] In contrast, the later-generated tryptic motifs were enriched in charged,

low specificity residues at P1' and P2' (Figure 2d). Similarly, when we extract motifs from miscleaved lysines and arginines from depleted HEK peptides we found that arginine cleavage sites were more enriched in the iceLogo (Figure 2c). This result implies that fast arginine sites were cleaved and depleted in our DigDeAPr strategy, while slow lysine sites remained missed cleaved and were not depleted. These conclusions are consistent with previous[23] and recent[8, 9] studies to improve lysine cleavage efficiency with tryptic digestions using Lys-C.

Supplementary Note 5: Further mechanistic insights from the analysis of yeast. Our yeast data provides an excellent means to further interrogate the DigDeAPr mechanism since both protein abundance[1] and identified "proteotypic" peptides[2] are well-characterized in yeast. Early- and later-generated peptides were correlated to both relative protein abundance measurements with spectral counts (Supplementary Figure 5a) and absolute protein abundance measurements by western blotting (Supplementary Figure 5b). We notably observed abundance-based trends with both spectral counting and absolute protein copies per cell. To further validate abundance-based depletion and enrichment trends observed with yeast cells, we also analyzed our HEK data in the same fashion. When early- and later-generated HEK peptides are considered in the context of protein abundance (based on spectral counts), we observe a similar abundance-based identification trend (Supplementary Figure 5c). These results would indicate that the analysis of a population of later-generated peptides contributes more to the identification and quantification of low abundance proteins than the analysis of a population of early-generated peptides.

We initially noticed that fast, tryptic cleavage motifs from HEK peptides (Figure 2b) also represent peptide motifs that should have a greater electrospray ionization (ESI) efficiency, due to their hydrophobicity and basicity. Thus, the peptides with these hydrophobic motifs would also be expected to have a greater mass spectrometer (MS) signal response.[24] Conversely, slow, missed cleavage sites, represented by acidic HEK peptide motifs in our analyses (Figure 2d) would tend to have a lower ESI efficiency and MS response.[25] A growing area of proteomic research is the characterization of "proteotypic" peptides which can be used as robust representations of protein identity and abundance.[2] We wondered if there may be a correlation to the robustly and reproducibly identified peptides, deemed "proteotypic", and the abundant, early-generated peptides that we depleted. We performed a Venn comparison of early- and later-generated peptides from our yeast analyses to proteotypic yeast peptides from MudPIT experiments analyzed by Mallick *et al*.[2] The Venn diagram (Supplementary Figure 7b) illustrates that there is indeed a high overlap of early-generated peptides and proteotypic peptides. We found that 58% of proteotypic peptides were considered early-generated, only 27% were later-generated, and about 14% of the proteotypic peptides were not found in either our control or DigDeAPr experiments. When we extracted motifs from the tryptic ends of proteotypic peptides (Supplementary Figure 7d) for comparison to our depleted, early-generated peptide motif (Supplementary Figure 7c), we found strikingly similar motifs: the four to five most represented amino acids at each position between the motifs are essentially the same. Furthermore, the extraction of motifs from the tryptic ends of yeast peptides that have been depleted based on spectral counts again show similar motifs to proteotypic peptides (Supplementary Figure

7e and f). These results indicate that, although proteotypic peptides are the most robust identifiers of a protein, they may also contribute the most to proteolytic background that inhibits the identification and coverage of low abundance proteins. Through depletion of these peptides, improvements to low abundance protein identification and quantification are realized.

Supplementary Note 6: Correlation of HEK peptide depletion or enrichment to estimated HEK protein abundance. The experimental results and kinetic derivations herein suggest the importance of a peptide-centric consideration of the DigDeAPr mechanism. We hypothesize that the tryptic digestion and LC-MS/MS pipeline may both contribute to the abundance-based effects observed in our DigDeAPr data set. This is partially illustrated by the effects of protease specificity on spectral counting-based quantitation.[10] Although we are using the same protease, trypsin, throughout our experiments, since we are sampling peptides with different tryptic specificities (V/K) in control and DigDeAPr runs, similar spectral counting-based quantitation effects may be observed. Thus, we performed further analyses using spectral counting methods to attempt to uncover any trends. An overlaid representation of peptide spectral count changes with protein spectral count changes from DigDeAPr (Supplementary Figure 6a) illustrates that both follow a similar trend. This is not unexpected since peptide spectral counts are summed to quantify proteins with spectral counts.[26] Thus, the abundance-based trend observed at the protein level may be due to the depletion or enrichment of just a few peptides. In fact, when we correlate the single most enriched peptide for enriched proteins (both measured by spectral counts) we observe an obvious trend (Supplementary Figure 6b, upper right quadrant). When the same correlation is

performed for the single most depleted peptide for depleted proteins, a similar trend is observed (Supplementary Figure 6b, lower left quadrant). Fitting this data to a line illustrates that changes in a single peptide's spectral count can account for ~ 30% (slope from Supplementary Figure 6b) of protein spectral count changes for 57% (R^2 from Supplementary Figure 6b) of the proteins. Thus depletion of high spectral count peptides can obviously also be interpreted as depletion of high spectral count, abundant proteins. However, since only 30% of the protein spectral count changes can be explained by the most changed peptide, other peptides with the protein may also contribute or other mechanisms entirely may be at play.

METHODS

Quantitative characterization of early- and later-generated peptides. Label-free chromatographic peak areas were extracted for both yeast and HEK cell data using Census.[27] Briefly, MS1 precursor isotope envelopes were extracted for identified peptides using a 30 ppm window and integrated over the chromatographic timescale. The exact same peptide sequences of different charge states were extracted and compared separately. Since peptides of the same charge state can be sampled multiple times during MudPIT, the peptide match with the highest XCorr, and presumably the highest signal, was extracted and integrated for comparison between separate MudPIT runs. Peptides with a $\log_2$ (DigDeAPr/Control) ratios $\leq$ -1 were considered early-generated while peptides with $\log_2$ (DigDeAPr/Control) ratios $\geq$ 1 were considered later-generated.

Quantitative characterization of tryptic motifs. Our previous database search considered an unlimited number of internal missed cleavages for each peptide candidate up to 6 kDa in length. Identified peptides were aligned to tryptic or missed cleaved lysine and arginine residues with Motif-x,[28, 29] then represented as motifs with iceLogo.[30] Positive data sets for iceLogo analyses were aligned tryptic ends of HEK and yeast peptides on depleted, early-generated peptides (considered fast cleavage sites) and enriched, later-generated peptides (considered slow cleavage sites). Missed cleavage sites within depleted, early-generated peptides were also considered fast cleavage sites. Peptides with $-1 < \log_2$ (DigDeAPr/Control) ratios < 1 were considered unchanged and used as the negative set of aligned sites for tryptic and missed cleavage motif extraction for HEK peptides. The regional-sampled UniProt yeast protein database was used as the negative set of sites for yeast peptide motif analyses.

REFERENCES

1. Ghaemmaghami, S. et al. *Nature* **425**, 737-741 (2003).
2. Mallick, P. et al. *Nature biotechnology* **25**, 125-131 (2007).
3. MacCoss, M.J. et al. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7900-7905 (2002).
4. Choudhary, G., Wu, S.L., Shieh, P. & Hancock, W.S. *Journal of proteome research* **2**, 59-67 (2003).
5. Swaney, D.L., Wenger, C.D. & Coon, J.J. *Journal of proteome research* **9**, 1323-1329 (2010).
6. Tran, B.Q. et al. *Journal of proteome research* **10**, 800-811 (2011).
7. Bian, Y. et al. *Journal of proteome research* **11**, 2828-2837 (2012).
8. Wisniewski, J.R. & Mann, M. *Analytical chemistry* **84**, 2631-2637 (2012).
9. Glatter, T. et al. *Journal of proteome research* **11**, 5145-5156 (2012).
10. Peng, M. et al. *Nature methods* **9**, 524-525 (2012).
11. Keil, B. Specificity of proteolysis. (Springer-Verlag, 1992).
12. Siepen, J.A., Keevil, E.J., Knight, D. & Hubbard, S.J. *Journal of proteome research* **6**, 399-408 (2007).

13. Fannes, T. et al. *Journal of proteome research* **12**, 2253-2259 (2013).
14. Markus, G., McClintock, D.K. & Castellani, B.A. *The Journal of biological chemistry* **242**, 4395-4401 (1967).
15. Caprioli, R.M., Fan, T. & Cottrell, J.S. *Analytical chemistry* **58**, 2949-2954 (1986).
16. Agard, N.J. et al. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1913-1918 (2012).
17. Fersht, A. Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding. (W.H. Freeman, New York; 1999).
18. Fonslow, B.R. et al. *Nature methods* **10**, 54-56 (2013).
19. Bakalarski, C.E. et al. *Journal of proteome research* **7**, 4756-4765 (2008).
20. Schwanhausser, B. et al. *Nature* **473**, 337-342 (2011).
21. Ning, K., Fermin, D. & Nesvizhskii, A.I. *Journal of proteome research* **11**, 2261-2271 (2012).
22. Wang, H., Alvarez, S. & Hicks, L.M. *Journal of proteome research* **11**, 487-501 (2012).
23. Washburn, M.P., Wolters, D. & Yates, J.R., 3rd *Nature biotechnology* **19**, 242-247 (2001).
24. Kulevich, S.E., Frey, B.L., Kreitinger, G. & Smith, L.M. *Analytical chemistry* **82**, 10135-10142 (2010).
25. McAlister, G.C. et al. *Analytical chemistry* **84**, 2875-2882 (2012).
26. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd *Analytical chemistry* **76**, 4193-4201 (2004).
27. Park, S.K., Venable, J.D., Xu, T. & Yates, J.R., 3rd *Nature methods* **5**, 319-322 (2008).
28. Schwartz, D. & Gygi, S.P. *Nature biotechnology* **23**, 1391-1398 (2005).
29. Chou, M.F. & Schwartz, D. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 13**, Unit 13 15-24 (2011).
30. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. *Nature methods* **6**, 786-787 (2009).