Addendum: Digestion and depletion of abundant proteins improves proteomic coverage

Bryan R Fonslow, Benjamin D Stein, Kristofor J Webb, Tao Xu, Jeong Choi, Sung Kyu Park & John R Yates III

Nature Methods **10**, 54–56 (2013); published online 18 November 2013; addendum published after print 27 February 2014

Recently we reported improved proteome coverage and quantitation metrics for low-abundance proteins within whole proteomes by implementing a digestion and depletion strategy (DigDeAPr) before a standard shotgun proteomic analysis. Our goal was to improve the detection of low-abundance proteins by reducing the proteolytic background of highly sampled peptides derived from high-abundance proteins. We rationalized that our gains in proteome coverage resulted from the selective digestion and removal of abundant proteins as peptides. Since the publication of the method, the mechanism by which our gains were achieved was challenged in a Correspondence by Ye *et al.*[1]. In response, we have reanalyzed our data in a peptide-centric manner and propose a refined kinetic mechanism consistent with established competitive substrate kinetics.

Through a simplified derivation beginning with a classical Michaelis-Menten competitive-substrate model and further quantitative analysis of our data, we provide a refined depletion mechanism that more accurately describes the complex mixtures we previously analyzed. Our revised qualitative expression describing depletion of early generated peptides from proximal fast tryptic cleavage sites with high specificity constants (*V*/*K*) (**Supplementary Note 1**) is illustrated by the following equation

$$\chi_{\mathrm{A,depleted}} = \frac{(\chi_{\mathrm{A}})_{t=0}\, e^{-\left(\frac{V}{K}\right)_{\mathrm{A}} t_{\mathrm{c}}}}{\sum (\chi_n)_{t=0}\, e^{-\left(\frac{V}{K}\right)_n t_{\mathrm{c}}}} - \frac{(\chi_{\mathrm{A}})_{t=0}\, e^{-\left(\frac{V}{K}\right)_{\mathrm{A}} t_{\mathrm{d}}}}{\sum (\chi_n)_{t=0}\, e^{-\left(\frac{V}{K}\right)_n t_{\mathrm{d}}}} \quad (1)$$

where $\chi_{\mathrm{A,depleted}}$ is the mole fraction of substrate A after complete (t_c) and depletion (t_d) digestion times expressed as mole fractions of total substrate cleavage sites. So expressed, tryptic sites have different specificity constants as well as abundances. Substrate cleavage results in the generation of two shorter polypeptides that can be subsequently cleaved into more substrates over time. The relative cleavage rates are governed by each site's relative specificity constant. From this perspective, we redefine the mechanism for depletion and enrichment of the DigDeAPr method. Early generated peptides, derived from fast substrate sites (i.e., those with high *V*/*K*) within ~100 amino acids of each other, are removed during our 10-kDa molecular-weight-cutoff spin-filter depletion step. The clearing of these early generated peptides before further digestion allows enrichment of peptides resulting from slower tryptic sites in the subsequent complete digestion step.

Using equation (1) we illustrate the expected adjustment in peptide abundance resulting from limited digestion and depletion (**Fig. 1a**) as driven by the relative cleavage-site specificity constants (*V*/*K*). When peptide abundance is considered between control and DigDeAPr runs, the expected trend is observed (**Fig. 1b** and **Supplementary Fig. 1**), a result consistent with our revised digestion and depletion theory. Notably, the use of tenfold more starting material and depletion of early generated peptides equalized the measured abundance of all peptides (**Supplementary Note 2** and

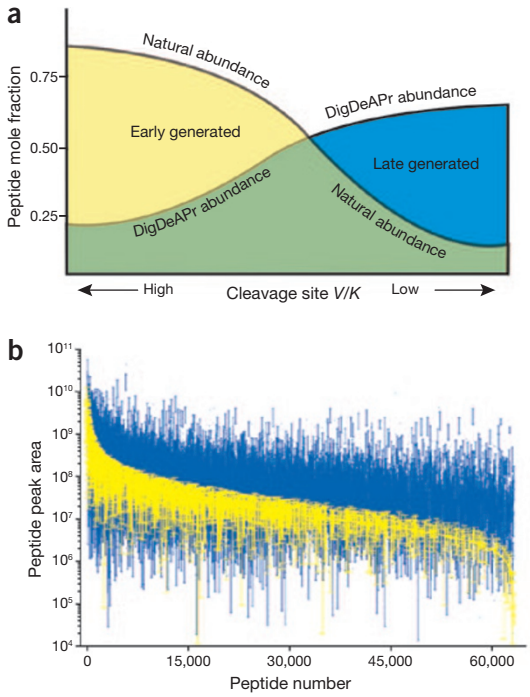

Figure 1 | Theoretical and empirical analysis of the DigDeAPr mechanism. (**a**) Schematic of our refined mechanism for digestion and depletion based on the cleavage-site specificity constant (*V*/*K*) for a given protease. The natural abundances of peptides from a complete protease digestion are adjusted by the use of ten times as much material and depletion of early generated peptides to enrich for late generated peptides with lower cleavage-site specificity constants. (**b**) Rank-abundance plot of peptide chromatographic peak areas from triplicate control (yellow) and DigDeAPr (blue) runs representing early and late generated peptides, respectively. Error bars, s.d.

Supplementary Figs. 2–4). Because peptide abundances are used to estimate protein abundance with shotgun proteomics[2–5], the equalization of peptides also equalizes the measurable abundance of proteins, as we found empirically in our initial analysis.

Our DigDeAPr runs provide a defined, limited digestion time point for consideration of the aforementioned kinetic efficiencies through analysis of early and late generated peptides and fast and slow tryptic cleavage sites (**Supplementary Note 3**). Early generated peptides should be depleted and have lower abundances after DigDeAPr when compared to control runs, whereas late generated peptides should be enriched and have higher abundances. Using label-free chromatographic peak-area ratios of peptides in both control and DigDeAPr runs, we quantified 13,628 and 13,112 peptides in human embryonic kidney (HEK) cells (**Fig. 2a**) and yeast cells, respectively, that were used to classify peptides as early or late generated by their relative ratios. Both distributions showed defined populations of peptides that were depleted ($\log_2$ ratio ≤ -1), unchanged ($-1 < \log_2$ ratio < 1) and enriched ($\log_2$ ratio ≥ 1). Focusing on the HEK peptide distribution, motif analysis of cleaved

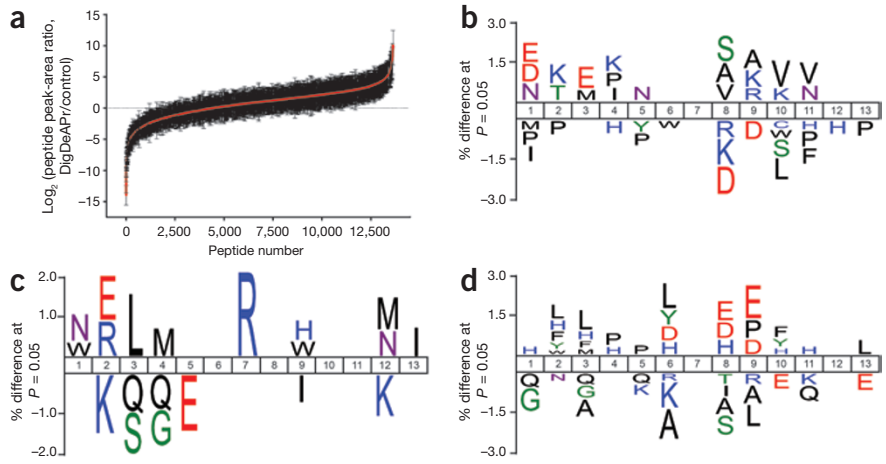

Figure 2 | Motif analysis to support the DigDeAPr mechanism. (**a**) Distribution of quantified HEK peptide ratios using label-free peak-area measurements (Online Methods). (**b**–**d**) Tryptic-site motif analysis using iceLogo[7] and peptide ratios from the analysis of HEK cells, categorized by peak-area ratios: depleted ($\log_2$ ratio ≤ −1) cleaved sites (n = 5,834) versus unchanged (−1 < $\log_2$ ratio < 1) cleaved sites (n = 11,885) (**b**), depleted missed cleavage sites (n = 2,846) versus unchanged missed cleavage sites (n = 5,438) (**c**), and enriched ($\log_2$ ≥ 1) cleaved sites (n = 22,239) versus unchanged cleaved sites (n = 11,885) (**d**).

(**Fig. 2b**) and missed cleaved (**Fig. 2c**) tryptic sites on depleted peptides confirmed that early generated peptides from proximal fast tryptic cleavage sites (<~100 amino acids apart) were selectively removed during the 10-kDa depletion step (**Supplementary Note 4**). Similarly, tryptic motifs of enriched, late generated peptides represent slow cleavage sites (**Fig. 2d**) that remained uncleaved within polypeptides of >10 kDa at the depletion time point. Thus, consideration of tryptic sites and peptides in the digestion and depletion mechanism is essential and illustrates the depletion and enrichment of peptides from fast and slow tryptic cleavage sites, respectively.

By considering these early and late generated peptides in our protein abundance analyses, we notably still observed an abundance-based depletion and enrichment trend in both yeast and HEK cells: higher-abundance proteins have more early generated peptides identified, and lower- abundance proteins have more late generated peptides identified (**Supplementary Fig. 5** and **Supplementary Note 5**). On the basis of these data and our understanding of peptide sampling in shotgun proteomics[2], we conclude that our gains originate from analysis of a different population of enriched, late generated peptides. That is, depletion of early generated peptides from high-abundance proteins removes enough proteolytic background to unmask and identify more late generated peptides from low-abundance proteins. Although we may not have explicitly depleted abundant proteins through digestion, in our reanalysis we found that depletion or enrichment of single peptides accounted for ~30% (1/slope = 0.298) of the observed protein abundance depletion or enrichment, respectively, explained by ~60% (coefficient of determination R^2 = 0.57) of the protein abundance measurements (**Supplementary Fig. 6** and **Supplementary Note 6**). Additionally, we found a notable overlap in depleted, early generated yeast peptides and 'proteotypic' yeast peptides (**Supplementary Fig. 7** and **Supplementary Note 5**). Although proteotypic peptides can be used to robustly identify and quantify many proteins, they can also act as proteolytic background for other less abundant or less sampled proteins and peptides[6]. Our results collectively indicate that depletion of highly sampled, abundant, easily identified, proteotypic peptides has a similar effect as depleting abundant proteins to improve identification and quantification of peptides from low-abundance proteins.

With our reexamined view of peptide abundance changes and their correlation to protein changes, we propose a refined mechanism by which our proteome coverage and quantitation gains are realized through digestion and depletion: depletion of early generated peptides and enrichment of late generated peptides equalizes measurable peptide abundances and unmasks less proteotypic peptides for improvements in low-abundance protein identification and quantification. We suggest that DigDeAPr should represent digestion and depletion of abundantly sampled peptides and proteins through enrichment of less easily digested and identifiable proteins and peptides. Nonetheless, the combination of tenfold more starting material with limited digestion and depletion remains a robust and straightforward method to remove the most easily and repeatedly detected peptides, clearing chromatographic, electrospray ionization and mass spectrometer space for improvements in identification coverage and quantification of low-abundance proteins. Our refined mechanistic analysis suggest that varying limited digestion times in combination with the use of other proteases with different site specificity constants (V/K) and different molecular-weight-cutoff filter sizes may hold the most potential to further improve coverage and quantitation of whole proteomes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth0314-345).

ACKNOWLEDGMENTS

This project was supported by the US National Center for Research Resources (5P41RR011823-17), National Institute of General Medical Sciences (8P41GM103533-17), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK074798), National Heart, Lung, and Blood Institute (RFP-NHLBI-HV-10-5) and National Institute of Mental Health (R01MH067880). We thank D. Schwartz for help with motif alignments and J. Moresco, J. Savas and A. Pinto for comments on the manuscript.

AUTHOR CONTRIBUTIONS

These additional analyses and derivations were performed by B.R.F. and Mark S. Hixon, respectively. M.S.H. (Department of Biological Sciences, Takeda California, San Diego, California, USA) provided valuable assistance with describing the enzyme kinetics of these complex mixtures. B.D.S., K.J.W., T.X., J.C., S.K.P. and J.R.Y. agree with the reanalysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ye, M., Pan, Y., Cheng, K. & Zou, H. *Nat. Methods* **11**, 221–222 (2014).
2. Liu, H., Sadygov, R.G. & Yates, J.R. III. *Anal. Chem.* **76**, 4193–4201 (2004).
3. Zybailov, B. *et al. J. Proteome Res.* **5**, 2339–2347 (2006).
4. Griffin, N.M. *et al. Nat. Biotechnol.* **28**, 83–89 (2010).
5. Schwanhäusser, B. *et al. Nature* **473**, 337–342 (2011).
6. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
7. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. *Nat. Methods* **6**, 786–787 (2009).

ONLINE METHODS

Quantitative characterization of early and late generated peptides. Label-free chromatographic peak areas (**Supplementary Data**) were extracted for both yeast and HEK cell data using Census[8]. Briefly, MS1 precursor isotope envelopes were extracted for identified peptides using a 30-p.p.m. window and integrated over the chromatographic timescale. The same peptide sequences of different charge states were extracted and compared separately. Because 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 ≤–1 were considered early generated, whereas peptides with $\log_2$(DigDeAPr/control) ratios ≥1 were considered late 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[9,10] and then represented as motifs with iceLogo[7]. 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, late generated peptides (considered slow cleavage sites). Missed cleavage sites within depleted, early generated peptides were also considered fast cleavage sites. Peptides with $\log_2$(DigDeAPr/control) ratios in the interval (–1, 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.

8. Park, S.K., Venable, J.D., Xu, T. & Yates, J.R. III. *Nat. Methods* **5**, 319–322 (2008).
9. Schwartz, D. & Gygi, S.P. *Nat. Biotechnol.* **23**, 1391–1398 (2005).
10. Chou, M.F. & Schwartz, D. *Curr. Protoc. Bioinformatics* **35**, 13.15 (2011).