Published in final edited form as:

J Proteome Res. 2010 March 5; 9(3): 1323–1329. doi:10.1021/pr900863u.

The value of using multiple proteases for large-scale mass spectrometry-based proteomics

Danielle L. Swaney¹, Craig D. Wenger¹, and Joshua J. Coon^{1,2,*}

¹Department of Chemistry, University of Wisconsin, Madison, WI 53706

²Department of Biomolecular Chemistry, University of Wisconsin, Madison, WI 53706

Abstract

Large-scale protein sequencing methods rely on enzymatic digestion of complex protein mixtures to generate a collection of peptides for mass spectrometric analysis. Here we examine the use of multiple proteases (trypsin, LysC, ArgC, AspN, and GluC) to improve both protein identification and characterization in the model organism *Saccharomyces cerevisiae*. Using a data-dependent, decision tree-based algorithm to tailor MS2 fragmentation method to peptide precursor, we identified 92,095 unique peptides (609,665 total) mapping to 3,908 proteins at a 1% false discovery rate (FDR). These results were a significant improvement upon data from a single protease digest (trypsin) – 27,822 unique peptides corresponding to 3,313 proteins. The additional 595 protein identifications were mainly from those at low abundances (i.e., < 1,000 copies/cell); sequence coverage for these proteins was likewise improved nearly 3-fold. We demonstrate that large portions of the proteome are simply inaccessible following digestion with a single protease and that multiple proteases, rather than technical replicates, provide a direct route to increase both protein identifications and proteome sequence coverage.

Keywords

Proteomics; Mass Spectrometry; Model Organisms

INTRODUCTION

Protein sequencing technologies have experienced rapid development over the past decade. In 2001, Washburn and co-workers described innovative peptide handling and fractionation methodology that enabled the identification of 1,484 proteins from yeast (*Saccharomyces cerevisiae*).¹ Since that time advances in peptide separations, mass spectrometry instrumentation, and informatics have enabled the identification of 4,621 proteins in this model organism that contains only 5,884 genes.^{1–4} With this success we now turn our attention from proteome identification to characterization. More specifically, the 5,884 yeast genes code for 2,916,123 non-redundant amino acids; however, only 889,216 of these have been observed by mass spectrometry-based proteomic analyses. Complete proteome characterization demands

*to whom correspondence should be addressed: jcoon@chem.wisc.edu.

AUTHOR CONTRIBUTIONS

DLS and JJC designed research;
DLS and CDW performed research;
DLS and JJC wrote the paper.

Competing Interest Statement. None.

Supporting Information Available: Supplementary Table 1, Data Set 1, and 2 are available free at <http://pubs.acs.org>.

the observation of each of these amino acids. Such coverage would allow one to comprehensively localize post-translational modifications (PTMs), differentiate homologous proteins, and detect post-transcriptional editing events.

For all the innovation that has enabled the identification of nearly every protein expressed in yeast, certain aspects of the method have not changed. Prominent among these is the near exclusive use of the protease trypsin to generate peptide fragments. Yeast tryptic peptides average 8.4 amino acids in length and contain a basic residue (Arg or Lys) on the C-terminus. When protonated and in the gas-phase, these peptide cations are ideal for sequencing by collisional activation tandem MS (*i.e.*, CAD MS²).^{5, 6} In general, peptides having low charge states (z) and high mass-to-charge ratios (m/z) are best sequenced *via* CAD, explaining the selection of trypsin.^{7, 8} A side-effect of trypsin digestion, however, is that the majority of generated peptides are very small ($56\% \leq 6$ residues) – too small for mass spectrometry-based sequencing. **Figure 1**, panel A, displays the theoretical length distribution of peptides following *in silico* digestion with trypsin. Superposed on these data is the distribution of identified tryptic peptides in the five large-scale yeast mass spectrometry-based proteomic experiments, including the data presented here (*vide infra*).¹⁻⁴ These data exhibit an obvious mismatch between optimal peptide length, for successful mass spectrometry-based sequence identification, and the *in silico* tryptic peptide distribution. Note 97% of all peptides identified in these collective works fall within a range of 7–35 residues.

Efforts to increase whole proteome coverage have historically focused on increasingly rigorous fractionation of complex tryptic peptide mixtures prior to mass spectrometric analysis. Figure 1 reveals that no matter the extent of fractionation, large segments of the proteome are sequestered and are simply not detectable as their primary sequence is incompatible with the applied technology (*e.g.*, ≤ 6 residues). A straightforward method to increase proteome coverage is to shift the distribution of peptides to more closely resemble those experimentally observed, by use of multiple proteases. MacCoss *et al.* and others recognized this several years ago and demonstrated a benefit using non-specific proteases.⁹⁻¹³ Non-specific proteases, however, can result in decreased reproducibility, increased sample complexity, and complicate quantification efforts. Other, more targeted experiments (<300 proteins), have utilized multiple proteases with good success.¹⁴⁻¹⁶

In silico calculations of peptide length distribution, following digestion of the yeast proteome with four commercially available proteases with specific cleavage chemistry, are shown in Figure 1A (LysC, AspN, GluC, and ArgC, panel A). We note that for these proteases, like trypsin, the most frequently generated peptide length is a single amino acid. To predict the collective impact of such an approach we independently considered each amino acid in the yeast proteome and ranked the sizes of the five peptides that contained it for each of the five proteases. In each instance we retained the peptide with the length that was most frequently observed in the experimental distribution (Fig. 1B). This confirms that, following digestion with five proteases, nearly all amino acids in this proteome (94.8%) are contained in at least one peptide of suitable length (7 – 35 residues) for mass spectrometry sequencing technology.

A limitation of a multiple protease strategy is that many of the optimal length peptides do not necessarily have CAD-friendly sequences; *i.e.*, contain only one Arg or Lys located at the C-terminus. Specifically, peptides among the expanded cohort are likely to contain one or more internal basic residues, a situation that can lead to unassignable CAD MS² spectra.^{5, 6, 17} Electron-based fragmentation methods (*i.e.*, ECD or ETD), however, are more or less indifferent to the presence of additional internal basic sites.^{18, 19} Thus, we postulated that the combination of CAD and ETD with a multiple protease digestion strategy could sizably improve protein sequence coverage on a large-scale. Recent work of ours has demonstrated that CAD and ETD identify distinct peptide populations.⁸ Hence, we reasoned the diverse

population of peptides produced from a variety of enzymes could be effectively countered by the combined use of CAD and ETD in a decision tree-driven fashion. In such a strategy the mass spectrometer system automatically determines which method is most likely to produce a sequence assignment based on precursor charge (z) and mass-to-charge ratio (m/z).⁸ Here we evaluate the use of a multiple protease strategy to improve our ability to comprehensively characterize entire proteomes.

EXPERIMENTAL PROCEDURES

Cell culture and protein harvesting

Wild-type *Saccharomyces cerevisiae* was grown in rich media to an OD₆₀₀ of 0.97 and centrifuged for 10 min at 4 °C. The pellet was washed twice with sterile water, and centrifuged at 3,000 rpm for 5 min. A volume of lysis buffer of approximately three times the cell pellet volume was added. The buffer contained 8 M urea, 75 mM NaCl, 50 mM Tris pH 8, 10 mM sodium pyrophosphate, complete mini EDTA-free protease inhibitor (Roche Diagnostics, Indianapolis, IN) and phosSTOP phosphatase inhibitor (Roche Diagnostics). The sample was French-pressed 3 times and centrifuged for 15 min at 14,000 rpm at 4 °C, and the protein containing supernatant stored at –80 °C.

Digestion

Cysteine residues were reduced and alkylated by incubation in 2.5 mM DTT for 25 min at 60 °C followed by incubation in 7 mM iodoacetamide in the dark at room temperature for 30 min. Alkylation was capped by incubation in 2.5 mM DTT for 15 min at room temperature. All digestions were performed under optimized conditions for each protease respectively. A 4 mg aliquot of protein was digested overnight at 37 °C after addition of CaCl₂ to 1 mM, 7M urea, and 40 µg of endoproteinase LysC (Princeton Separations, Adelphia, NJ). A 2 mg aliquot of protein was digested in 2 M urea, 50 mM Tris, 25 mM ammonium bicarbonate at room temperature overnight with 25 µg of GluC (Roche Diagnostics, Indianapolis, IN). A 2 mg aliquot of protein was digested in 4 M urea, 100 mM Tris, 10 mM CaCl₂, 0.5 mM EDTA, 5 mM DTT at 37 °C overnight with 10 µg of ArgC (Roche Diagnostics, Indianapolis, IN). A 2 mg aliquot of protein was digested in 1.6 M urea, 10 mM Tris at 37 °C over night with 6 µg of Asp-N (Roche Diagnostics, Indianapolis, IN). Finally, a 2 mg aliquot of protein was digested in 1.5 M urea, 50 mM Tris, and 10 mM CaCl₂ at 37 °C overnight with 20 µg of trypsin (Promega, Madison, WI). Each digest was quenched by the addition of TFA to a final concentration of 0.5%, desalted on a 100 mg tC₁₈ SepPak cartridge (Waters, Milford, MA), and the eluent lyophilized.

Fractionation

SCX fractionation of all samples was performed as previously described^{8, 20}. Each peptide digest was dissolved in 500 µL of SCX buffer A (5 mM KH₂PO₄, 30% acetonitrile, pH 2.65) and loaded onto a polysulfoethyl aspartamide column (9.4-×200 mm, PolyLC, Columbia, MD) connected to a Surveyor LC quaternary pump (Thermo Electron, San Jose, CA) running at 3.0 mL/min. Peptides were detected *via* a PDA detector (Thermo Electron, San Jose, CA). A total of 12 fractions were collected in 4 min intervals for each peptide digest over the course of the SCX gradient. For all but the tryptic digest the following gradient was used: 2 min of isocratic buffer A, followed by a linear gradient of 0–15% buffer B from 2 to 5 min (5 mM KH₂PO₄, 30% acetonitrile, 350 mM KCl, pH 2.65), followed by a linear gradient of 15–100% buffer B from 5 min to 35 min. Buffer B was held at 100% for 6 min. For the tryptic digest the following gradient was used: 2 min of isocratic buffer A, followed by a linear gradient of 0–10% buffer B from 2 to 5 min (5 mM KH₂PO₄, 30% acetonitrile, 350 mM KCl, pH 2.65), followed by a linear gradient of 10–60% buffer B from 5 min to 35 min. Buffer B was ramped up to 100% over the next 6 min. For all digests there was then a 7 min transition from buffer B to 100%

buffer C (50 mM KH_2PO_4 , 500 mM KCl, pH > 7.5). Finally, Buffer C and Buffer D (nanopure water) were used to wash the column. Each fraction was lyophilized, and desalted on 50 mg tC_{18} SepPak cartridges (Waters, Milford, MA). Desalted eluates were lyophilized, resuspended in 0.2% formic acid, and stored at -20°C .

nanoHPLC

A Waters nanoAquity HPLC and autosampler were used to load and chromatographically separate SCX fractions. Samples were loaded onto a pre-column, and separated on a 50 μm i.d. analytical columns packed to 12 cm, as previously described.²¹ Sample loading amounts were adjusted for each fraction such that similar MS^1 base peak intensity was obtained. Initial CAD-only and ETD-only analyses were performed using a 40 min linear gradient of 1.4% to 49% acetonitrile in 0.2% formic acid. All decision-tree acquisitions were performed using a 120 min linear gradient from 4% to 30% acetonitrile in 0.2% formic acid.

Mass Spectrometry

All experiments were performed on an ETD-enabled hybrid linear ion trap-orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany).^{22, 23} nanoHPLC eluates were directly sampled *via* an integrated electrospray emitter operating a 2.3 kV. Initial experiments consisted of MS^1 analysis in the orbitrap mass analyzer followed by six data-dependent MS^2 events with mass analysis in the ion trap. The type of dissociation in each MS^2 event was either CAD or ETD for all six MS^2 events. For triplicate experiments utilizing decision tree-based MS^2 acquisition, orbitrap MS^1 analysis was followed by eight data-dependent MS^2 events utilizing either ETD or CAD interrogation.⁸ For all experiments a target value of 10,000 charges was used for QIT MS^2 AGC, precursors were dynamically excluded for 40 s, and only peptides with assigned charge states of two or greater were selected for MS^2 interrogation. All data files are publically available at ProteomeCommons.org (hash to be inserted upon manuscript acceptance).

Database searching

Peak lists were generated using DTA Generator (<http://www.chem.wisc.edu/~coon/software.html>) using an absolute fragment intensity of zero. For ETD spectra the precursor, charge-reduced precursor ions, and peaks corresponding to neutral losses were removed.²⁴ The processed spectra were then searched against a concatenated target-decoy version of the *Saccharomyces* Genome Database (<http://www.yeastgenome.org>, downloaded 02/04/2009) using OMSSA (Open Mass Spectrometry Search Algorithm version 2.1.4).^{25, 26} The search algorithm parameters were set to consider static modifications of +57.021464 Da on cysteine residues (carbamidomethylation), differential modifications of +15.994915 Da on methionine residues (oxidation) and +42.01 Da on the protein N-terminus (acetylation), a precursor mass tolerance of ± 4.0 Da, a fragment ion mass tolerance of ± 0.5 Da, and a maximum of 3 missed cleavages. Tryptic peptides were searched with Arg and Lys cleavage specificity, GluC with Glu, AspN with Asp, ArgC with Arg, and LysC with Lys cleavage specificity. An in-house program was used to trim all identifications by identification score and precursor error so that the entire data set for each protease had a FDR of 1% and were within ± 7 ppm of the theoretical precursor m/z . Next, peptides were assigned to protein groups, such that the smallest number of proteins were represented. These protein groups were assigned a p-score and filtered to a 1% FDR at the protein level. Finally, the peptide list was reduced to represent only peptides from proteins identified at a 1% FDR.^{25, 27} The P-score was calculated by multiplying the identification scores of all unique peptides within a given protein group.

RESULTS AND DISCUSSION

Experimental validation

Five commercially available proteases with high specificity were selected for comparison, and digest conditions were independently optimized for each protease. After defining optimal digest conditions, aliquots of a yeast whole cell lysate were digested overnight, separately, with either trypsin, LysC, ArgC, GluC, or AspN. Peptides resulting from each digest were separated into 12 fractions *via* strong cation exchange (SCX) chromatography.²⁰ Each of the 60 fractions was analyzed in quadruplicate *via* nanoflow reversed-phase chromatography wherein the effluent was directed into an ETD-enabled linear ion trap-orbitrap hybrid mass spectrometer (nHPLC-MS²) where dissociation was accomplished either with CAD or ETD (two analyses with each). The orbitrap was used for MS¹ scans, while all MS² scans were executed in the ion trap. The goal of these experiments was to determine the optimal decision tree branch points for peptides from each protease. The resulting tandem mass spectra were searched against the *Saccharomyces cerevisiae* genome database (<http://www.yeastgenome.org>) using OMSSA.²⁶ Spectral matches were then filtered to a 1% false discovery rate (FDR) at spectral level.²⁵ From these data the probability of peptide identification was calculated as a function of precursor z and m/z (data not shown).⁸ Surprisingly, the m/z branch points of the decision tree were the same for peptides from all proteases.

Each of the 60 fractions was further analyzed in triplicate by nHPLC-MS² using the decision tree-driven data dependent algorithm. This algorithm applies the method of fragmentation (CAD or ETD) with the highest probability of generating a successful peptide identification for every precursor selected for MS² in an automated fashion. After database searching, as described above, spectral matches were filtered to a 1% FDR at the protein level (Fig. 2). In total 2.6×10^6 tandem mass spectra, mapping to 92,095 unique peptides (609,665 total) and 3,908 proteins at a 1% FDR, were acquired in the decision tree-driven acquisitions. These results are displayed in Table 1. A complete list of identified peptides and proteins can be found online (Supplementary Data Set 1 and 2 respectively). The trypsin dataset comprised the largest number of unique peptide identifications (27,822), followed by AspN (21,654), LysC (20,619), GluC (17,968), and ArgC (12,452). Collectively, these peptides encompass 742,312 non-redundant amino acids. Panel A of Figure 3 displays the overlap between data resulting from trypsin digestion as compared to the four other proteases. Use of the additional proteases more than doubled the amino acid coverage. Peptide identifications across the five protease datasets roughly correlate with average *in silico* peptide length. Specifically, digestion with proteases that generate the most peptides (*i.e.*, trypsin, 8.4 residues), which are in turn shorter on average, resulted in more peptide identifications than those that produced fewer (*i.e.*, ArgC, 21.4 residues). These data are plotted in Figure 4. Despite these differences the experimental distribution of peptide lengths identified following digestion with each protease was similar.

Trypsin was the only protease for which more precursors were selected for CAD MS² - illustrating why trypsin has traditionally been the protease of choice. The use of proteases other than trypsin produced peptides that were less favorable for CAD; however, the heterogeneity was countered by the combined use of CAD and ETD in a data-dependent decision tree-driven analysis. We note ETD was reasonably effective at sequencing tryptic peptides - contributing ~ 44% (12,356) of the identifications. Of the 92,095 unique peptide identifications resulting from use of all five proteases, almost 60% (53,920) were the result of ETD fragmentation; demonstrating that no matter which protease is used, the joint use of CAD and ETD is beneficial.

Next we examined the number of identified proteins and proteome sequence coverage - two critical figures of merit - to determine (1) the viability of using multiple proteases and (2) whether multiple replicates of a single protease sample would provide similar results. First,

triplicate analysis of any single protease sample resulted in an average of 3,010 protein identifications ($\sigma = 251$) with the tryptic dataset topping the list at 3,313 (Table 1). Summation of protein identifications from all five datasets increased this number by 595 proteins to 3,908 (18% increase) over trypsin alone (**Fig. 3**, panel B). The mean number of non-redundant amino acids sequenced by each of the five experiments was 273,244 ($\sigma = 60,456$). Again, the trypsin dataset topped the list with 346,510 amino acids; however, summation of all five datasets resulted in a considerable increase of 395,802 additional amino acids for a total of 742,312 (Table 1 and **Fig. 3**, panel A). Panel C of Figure 3 displays the impact of including additional proteases on the number of protein identifications and sequence coverage – a 172% increase in sequence coverage from one protease to five. As additional proteases are used, the mean number of proteins identified increases by an average of 6.9%, whilst the mean proteome coverage increases by a sizeable 30.0%. The greatest contributions are made by the addition of data from a second protease – protein identifications rose by 15.8% and average sequence coverage by 64.9%.

We considered similar results might be attainable by simply performing more technical replicates of a single protease sample. **Figure 3**, panel D, compares the percent increase in sequence coverage following a single technical replicate of multiple protease digests versus multiple technical replicates of a single protease digest. After three single technical replicates of three separate protease digests a 116% boost in sequence coverage is attained; only a 24% increase was observed following technical replicates of a single protease digest. The addition of data from a third protease contributed 5.7% more proteins and 27.5% more amino acids – a significant improvement over performing a third technical replicate of a single protease sample (3.2% boost in identifications and 4.7% increase in unique amino acids). We note other studies have reported similar diminishing returns for numerous technical replicates.²⁸ Thus, multiple proteases can enable access to segments of the proteome that are invisible upon digestion with a single protease on a large-scale.

Figure 5 plots protein identifications and proteome sequence coverage as a function of protein abundance.^{29, 30} These data demonstrate that high abundance proteins (> 100,000 copies per cell) are readily identified by use of a single protease, but that only about 30% of lower abundance proteins (< 1,000 copies per cell) are detected following digestion with a single protease. Cumulative protein identifications for all protease digests, shown in solid black bars (**Fig. 5**, panel A), are nearly double that for a single protease digest. Note the additional identifications are mainly from those at the lowest abundance. Such proteins may present only a single opportunity for identification; thus, if a peptide does not possess favorable characteristics for MS² sequencing, it will not be identified – as evidenced by the failure of multiple (*i.e.*, more than three) technical replicates of a single protease digest to significantly increase the overall identification rate. This effect, however, can be alleviated by use of multiple proteases, as different peptides will be produced, increasing the chances for successful identification. **Figure 5**, panel B, illustrates that use of multiple proteases has an even greater impact on proteome sequence coverage (*i.e.*, the total number of non-redundant amino acids sequenced). Increased sequence coverage was observed across all ranges of protein abundance following combination of five protease data. For very low abundance proteins (< 100 copies/cell) only 2.6% of amino acids were sequenced following analysis of the trypsin dataset – this number was increased nearly threefold to 7.1% when data from all proteases were compiled. High abundance proteins (> 100,000) had 75.6% of their amino acids sequenced upon analysis of all five protease digests (Figure 5B).

Identification of dubious open reading frames

The *Saccharomyces* genome database (SGD) contains 812 dubious proteins. Such proteins provide a means to evaluate the FDR estimate established by the decoy database strategy. Our

study identified only three of the 812 dubious proteins, suggesting that our maximum protein FDR of 1% provides a conservative estimate (Supplementary Table 1). YBR126W-A, one of these questionable proteins, was identified by six peptides covering 64.7% of the protein sequence (44 of 68 amino acids) in our work, and by four peptides in another study.² Further, this is the only protein in the SGD database to which these peptides can be matched, providing additional evidence that this open reading frame (ORF) does indeed encode a protein. The other two dubious ORFs identified here were YEL068C (by one peptide) and YAR075W (17 peptides, 101 of 157 amino acids). We note that of the 17 peptides identified as corresponding to YAR075W, only one is a unique match to that protein, while the remaining peptides could be explained by a combination of two verified proteins (YHR216W and YLR432W). The rules of parsimony, however, dictate the simplest explanation; hence, the dubious protein was selected to represent this group of peptides.²⁷

Implications for post-translational modification mapping

Here we have characterized the capability of a multiple protease approach to substantially increase proteome coverage – a figure of merit that is critical for complete proteome characterization, including PTMs.^{11, 31, 32} To gauge the large-scale PTM-mapping potential of this method, we counted the number of serine (S) and threonine (T) residues present in the yeast proteome (434,476) and then examined how many of these were identified among the various datasets. To generate the most inclusive dataset possible we expanded the data collected here to include four other large-scale yeast experiments on record. Three of those four works used only trypsin, while one of them, the 2008 study by Mann *et al.*, used both trypsin and the protease LysC.¹⁻⁴ We combined all of the tryptic peptide identifications with our own and pooled previously reported LysC sequences with those we generated from LysC, ArgC, AspN, and GluC. The tryptic dataset contained 92,023 S/T residues, while the non-tryptic dataset comprised 125,282. Concatenation of both datasets yielded coverage of 151,293 non-redundant S/T residues – a 64.4% improvement over the trypsin datasets, which were significantly larger in size than the non-tryptic data. This confirms previous reports demonstrating the utility of multiple proteases for PTM characterization.^{31, 32}

CONCLUSIONS

With theory and experiment we describe here the benefits that can be achieved by the use of multiple proteases in large-scale mass spectrometric-based protein sequencing. For the model organism yeast we observed a modest boost in protein identifications ~ 20% over use of a single protease, but a more than twofold improvement in proteome sequence coverage. Note the additional protein identifications were mainly derived from those occurring at the lowest abundances (< 1,000 copies per cell). Such results were not attainable by performing multiple technical replicates of a single protease digest. We rationalize these results by the fact that the MS-based analysis is best-suited for the detection of peptides of a particular size (~7–35 residues) and amino acid composition (i.e. a mobile proton for CAD, or a sufficient number of basic residues for ETD interrogation). Upon digestion with any single protease, certain portions of the proteome will be present in peptides that are simply not of suitable size for the conventional methodology. By changing cleavage specificity (*i.e.*, proteases), different portions of the proteome come into, or go out of, view of the MS technology. Additionally, the use of proteases with specific cleavage chemistry permits the use of the approach for quantitative studies.

As proteomic technologies continue to evolve emphasis is increasingly shifted from large-scale identification to comprehensive characterization – that is, the detection of every predicted amino acid in a given proteome. Such capabilities are critical for mapping PTMs, detecting SNPs, and identification of post-transcriptional editing. Here we have demonstrated that the

relatively small amount of available non-trypsin data for the organism yeast has already greatly expanded the number of observed S/T residues from 92,023 to 151,293. For all proteases other than trypsin, ETD identified more peptides than CAD. In total, 60% of the 92,095 unique peptide identifications resulted from ETD fragmentation. These results confirm that the chemical diversity of non-tryptic peptides is best countered by use of multiple dissociation methods in a decision tree-driven fashion and that less substantial gains will be observed if only one method is utilized. Peptide identifications across the five proteases roughly correlate with average *in silico* peptide length, indicating that more sizeable gains could be made by optimization of the separation methods for longer peptides.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the University of Wisconsin, the Beckman Foundation, and the National Institutes of Health (NIH) grants R01GM080148 (to J.J.C.) provided financial support for this work. D.L.S. acknowledges support from an NIH pre-doctoral traineeship – the Genomic Sciences Training Program, NIH 5T32HG002760.

References

1. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotech* 2001;19(3):242–247.
2. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008;455(7217):1251–1254. [PubMed: 18820680]
3. de Godoy L, Olsen J, de Souza G, Li G, Mortensen P, Mann M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biology* 2006;7(6):R50. [PubMed: 16784548]
4. Peng JM, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *Journal of Proteome Research* 2003;2(1):43–50. [PubMed: 12643542]
5. Dongre AR, Jones JL, Somogyi A, Wysocki VH. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *Journal of the American Chemical Society* 1996;118(35):8365–8374.
6. Huang YY, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry* 2005;77(18):5800–5813. [PubMed: 16159109]
7. Good DM, Wirtala M, McAlister GC, Coon JJ. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. *Mol Cell Proteomics* 2007;6(11):1942–1951. [PubMed: 17673454]
8. Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* 2008;5(11):959–964. [PubMed: 18931669]
9. Aebersold RH, Leavitt J, Saavedra RA, Hood LE, Kent SBH. INTERNAL AMINO-ACID SEQUENCE-ANALYSIS OF PROTEINS SEPARATED BY ONE-DIMENSIONAL OR TWO-DIMENSIONAL GEL-ELECTROPHORESIS AFTER INSITU PROTEASE DIGESTION ON NITROCELLULOSE. *Proc. Natl. Acad. Sci. U. S. A* 1987;84(20):6970–6974. [PubMed: 3313383]
10. Harper, RG.; Workman, SR.; Schuetzner, S.; Timperman, AT.; Sutton, JN. Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. Wiley-VCH Verlag GmbH; 2004. p. 1299-1306.
11. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, Clark JI, Yates JR. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U. S. A* 2002;99(12):7900–7905. [PubMed: 12060738]

12. Schlosser A, Vanselow JT, Kramer A. Mapping of Phosphorylation Sites by a Multi-Protease Approach with Specific Phosphopeptide Enrichment and NanoLC-MS/MS Analysis. *Analytical Chemistry* 2005;77(16):5243–5250. [PubMed: 16097765]
13. Wang B, Malik R, Nigg EA, Kořlner R. Evaluation of the Low-Specificity Protease Elastase for Large-Scale Phosphoproteome Analysis. *Analytical Chemistry* 2008;80(24):9526–9533. [PubMed: 19007248]
14. Biringer RG, Amato H, Harrington MG, Fonteh AN, Riggins JN, Huhmer AFR. Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. *Brief Funct Genomic Proteomic* 2006;5(2):144–153. [PubMed: 16772279]
15. Choudhary G, Wu S-L, Shieh P, Hancock WS. Multiple Enzymatic Digestion for Enhanced Sequence Coverage of Proteins in Complex Proteomic Mixtures Using Capillary LC with Ion Trap MS/MS. *Journal of Proteome Research* 2003;2(1):59–67. [PubMed: 12643544]
16. Elenitoba-Johnson KSJ, Crockett DK, Schumacher JA, Jenson SD, Coffin CM, Rockwood AL, Lim MS. Proteomic identification of oncogenic chromosomal translocation partners encoding chimeric anaplastic lymphoma kinase fusion proteins. *Proc. Natl. Acad. Sci. U. S. A* 2006;103(19):7402–7407. [PubMed: 16651537]
17. Kapp EA, Schutz F, Reid GE, Edes JS, Moritz RL, O'Hair RAJ, Speed TP, Simpson RJ. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry* 2003;75(22):6251–6264. [PubMed: 14616009]
18. Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A* 2004;101(26):9528–9533. [PubMed: 15210983]
19. Zubarev RA, Kelleher NL, McLafferty FW. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* 1998;120(13):3265–3266.
20. Villen J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(5):1488–1493. [PubMed: 17242355]
21. Ficarro SB, Zhang Y, Lu Y, Moghimi AR, Askenazi M, Hyatt E, Smith ED, Boyer L, Schlaeger TM, Luckey CJ, Marto JA. Improved Electrospray Ionization Efficiency Compensates for Diminished Chromatographic Resolution and Enables Proteomics Analysis of Tyrosine Signaling in Embryonic Stem Cells. *Analytical Chemistry* 2009;81(9):3440–3447. [PubMed: 19331382]
22. McAlister GC, Berggren WT, Griep-Raming J, Horning S, Makarov A, Phanstiel D, Stafford G, Swaney DL, Syka JEP, Zabrouskov V, Coon JJ. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *Journal of Proteome Research* 2008;7(8):3127–3136. [PubMed: 18613715]
23. McAlister GC, Phanstiel D, Good DM, Berggren WT, Coon JJ. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Analytical Chemistry* 2007;79(10):3525–3534. [PubMed: 17441688]
24. Good DM, Wenger CD, McAlister GC, Bai DL, Hunt DF, Coon JJ. Post-Acquisition ETD Spectral Processing for Increased Peptide Identifications. *Journal of the American Society for Mass Spectrometry* 2009;20(8):1435–1440. [PubMed: 19362853]
25. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007;4(3):207–214. [PubMed: 17327847]
26. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang XY, Shi WY, Bryant SH. Open mass spectrometry search algorithm. *Journal of Proteome Research* 2004;3(5):958–964. [PubMed: 15473683]
27. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Mol Cell Proteomics* 2005;4(10):1419–1440. [PubMed: 16009968]
28. Ham BM, Yang F, Jayachandran H, Jaitly N, Monroe ME, Gritsenko MA, Livesay EA, Zhao R, Purvine SO, Orton D, Adkins JN, Camp DG, Rossie S, Smith RD. The Influence of Sample Preparation and Replicate Analyses on HeLa Cell Phosphoproteome Coverage. *Journal of Proteome Research* 2008;7(6):2215–2221. [PubMed: 18412383]

29. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. Global analysis of protein expression in yeast. *Nature* 2003;425(6959):737–741. [PubMed: 14562106]
30. Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. Global analysis of protein localization in budding yeast. *Nature* 2003;425(6959):686–691. [PubMed: 14562095]
31. Gauci S, Helbig AO, Slijper M, Krijgsveld J, Heck AJR, Mohammed S. Lys-N and Trypsin Cover Complementary Parts of the Phosphoproteome in a Refined SCX-Based Approach. *Analytical Chemistry* 2009;81(11):4493–4501. [PubMed: 19413330]
32. Molina H, Horn DM, Tang N, Mathivanan S, Pandey A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *PNAS* 2007;104(7):2199–2204. [PubMed: 17287340]

Figure 1.

Plot of peptide length distribution for yeast proteome. Panel A displays peptide length profile for five proteases following an *in silico* digestion of the yeast proteome. Also shown is a plot of experimentally identified tryptic peptides – these peptides were drawn from five recent publications.¹⁻⁴ We independently considered each amino acid in the yeast proteome and ranked the sizes of the five peptides that contained it for each of the five proteases from panel A. In each instance we retained the peptide with the length that was most frequently observed in the experimental distribution. This best case distribution is plotted in Panel B and confirms that nearly all amino acids in the yeast proteome (94.8%) are contained in at least one peptide of suitable length for MS sequencing technology.

Figure 2.

Experimental workflow. Following isolation, proteins from *Saccharomyces cerevisiae* cells, were separated into aliquots and digested with one of the following proteases: trypsin, LysC, ArgC, GluC, ApsN. Each digest was independently fractionated *via* strong cation exchange, followed by reversed-phase nano HPLC- MS². The method of MS² was selected using a decision tree-driven approach. All data was then searched against the *Saccharomyces* Genome Database using OMSSA and filtered first to a 1% FDR at the peptide level, and finally to a 1% FDR at the protein level.

Figure 3.

Comparison of protein and non-redundant amino acid identifications. The overlap of non-redundant amino acid identifications (a) and proteins (b) between trypsin and the combined datasets from ArgC, AspN, GluC, and LysC. The number of identification unique to each group alone is displayed along with the percent overlap. The percent increase in proteins and non-redundant amino acids when comparing the mean of triplicate analyses of a single protease to the mean of any permutation of additional protease is shown in panel c. Panel d presents a comparison of single replicates of different protease vs. technical replicates of a single protease. Error bars represent the maximum and minimum percent increases observed and, in panel c, the protease combinations resulting in the maximum amino acid identifications are displayed above.

Figure 4.

Assessment of *in silico* and experimental peptide length. Panel a presents the average peptide length as calculated in silico. Panel b displays the number of unique peptide identifications vs. average *in silico* peptide length for each protease dataset. Finally, panel c shows the experimental distribution of peptide lengths resulting from cleavage with 5 different proteases.

Figure 5.

Evaluation of protein identifications and protein sequence coverage as a function of protein abundance. The viewgraph in panel a displays the percentage of proteins identified from an individual protease dataset and collectively (black bars). Panel b presents sequence coverage in a similar fashion.

Table 1

Summary of amino acid, peptide, and protein identifications.

Protease	Trypsin	ArgC	AspN	GluC	LysC	All
Unique peptides	27822	12,452	21,654	17,968	20,619	92,095
CAD	15466	3,518	9,267	7,331	7,807	38,175
ETD	12356	8,934	12,387	10,637	12,812	53,920
Total scans	538,175	540,674	514,607	507,278	524,764	2,625,498
Proteins	3,313	2,708	3,183	2,813	3,030	3,908
Percent of ORFs	56.3	46.0	54.1	47.8	51.5	66.4
Non-redundant amino acids	346,510	191,686	287,188	235,851	304,984	742,312
Non-redundant amino acid proteome coverage (percent)	11.9	6.6	9.8	8.1	10.5	25.5
Average protein sequence coverage (percent)	24.5	18.6	21.5	20.9	24.3	43.4