

Comprehensive Analysis of Protein Digestion Using Six Trypsins Reveals the Origin of Trypsin As a Significant Source of Variability in Proteomics

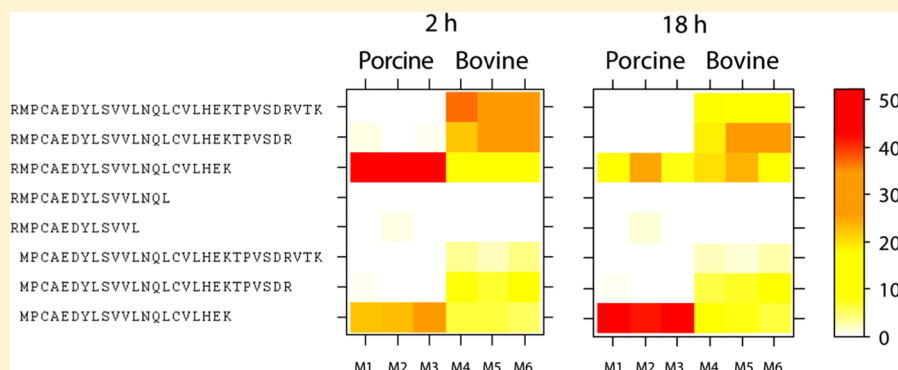
Scott J. Walmsley,[†] Paul A. Rudnick,[‡] Yuxue Liang,[‡] Qian Dong,[‡] Stephen E. Stein,^{*,‡} and Alexey I. Nesvizhskii^{*,†,§}

[†]Department of Pathology, University of Michigan, 4237 Medical Science I, 1301 Catherine Road, Ann Arbor, Michigan 48109, United States

[‡]National Institute of Standards Technology, 100 Bureau Drive, MS 8362, Gaithersburg, Maryland 20899, United States

[§]Department of Computational Medicine and Bioinformatics, University of Michigan, Room 2017 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, Michigan, United States

S Supporting Information



ABSTRACT: Trypsin is an endoprotease commonly used for sample preparation in proteomics experiments. Importantly, protein digestion is dependent on multiple factors, including the trypsin origin and digestion conditions. In-depth characterization of trypsin activity could lead to improved reliability of peptide detection and quantitation in both targeted and discovery proteomics studies. To this end, we assembled a data analysis pipeline and suite of visualization tools for quality control and comprehensive characterization of preanalytical variability in proteomics experiments. Using these tools, we evaluated six available proteomics-grade trypsins and their digestion of a single purified protein, human serum albumin (HSA). HSA was aliquoted and then digested for 2 or 18 h for each trypsin, and the resulting digests were desalted and analyzed in triplicate by reversed-phase liquid chromatography–tandem mass spectrometry. Peptides were identified and quantified using the NIST MSQC pipeline and a comprehensive HSA mass spectral library. We performed a statistical analysis of peptide abundances from different digests and further visualized the data using the principal component analysis and quantitative protein “sequence maps”. While the performance of individual trypsins across repeat digests was reproducible, significant differences were observed depending on the origin of the trypsin (i.e., bovine vs porcine). Bovine trypsins produced a higher number of peptides containing missed cleavages, whereas porcine trypsins produced more semitryptic peptides. In addition, many cleavage sites showed variable digestion kinetics patterns, evident from the comparison of peptide abundances in 2 h vs 18 h digests. Overall, this work illustrates effects of an often neglected source of variability in proteomics experiments: the origin of the trypsin.

KEYWORDS: proteomics, mass spectrometry, trypsin, digestion, endoprotease specificity, peptide abundance, variability, missed cleavages, label-free quantification, statistical analysis

INTRODUCTION

Mass spectrometry (MS)-based proteomics is a key technology used in biomedical research. Both targeted and discovery proteomics strategies are being increasingly used for characterization of peptides and proteins using samples derived from cells, tissues, or biological fluids such as plasma or urine.^{1,2} MS technology is also being increasingly applied in the analysis of biological pharmaceuticals, food allergen detection, and related

applications.^{3–5} All of these applications rely on the ability to identify and quantify biological molecules such as peptides and proteins in the analyzed samples with a high degree of accuracy, sensitivity, and reproducibility. At the same time, a typical quantitative proteomics workflow is a complex process

Received: June 26, 2013

Published: October 11, 2013

consisting of multiple preanalytical (sample processing) and analytical steps.⁶ It includes proteolytic digestion of proteins into peptides, separation of the resulting peptide mixtures using liquid chromatography (LC), identification and quantification of peptides by MS, and computational analysis of MS data. As a result, technical variability in proteomics experiments can be high. Understanding and measuring all major sources of variability is essential for the success of MS-based proteomics as a reliable measurement platform.^{7–10}

In a typical analysis, proteins are assayed by the detection of their corresponding fragments – peptides – produced by proteolytic digestion of proteins using an endoprotease.¹¹ The abundance of a peptide is estimated using the intensity of the peptide ion extracted from MS data¹² (or using fragment ion intensities in the case of targeted strategies such as selected reaction monitoring (SRM)¹³ or SWATH-MS¹⁴, with an additional (optional) normalization to the intensity of a spiked-in labeled reference peptide. Nondetection of peptide ions corresponding to high charge states of the selected peptide, unanticipated chemical modifications, incomplete digestion resulting in peptides containing missed cleavage sites, and run-to-run variation in the ionization processes such as in-source fragmentation of a parent peptide sequence all contribute to errors in the measurements of the peptide abundance. Total measurement variance has been studied, indicating that interlaboratory CVs for quantified peptides can be as high as >25%, albeit these are both laboratory- and target-sequence-dependent.^{7,9,10} More recent work developed statistical models to elucidate primary contributions toward variance while also showing the advantage of using heavy isotope-labeled reference material.^{15,16} These studies have shown that the proteolytic digestion is a primary source of error in the abundance measurement.^{7,17,18} Gaining a better understanding of proteolytic digestion and its contribution toward measurement variability in proteomics is the main motivation behind our work.

Trypsin is the most commonly used endoprotease for proteomic analysis, and several commercially available enzymes have been produced. Most commonly used are ‘proteomics-grade’ trypsins, where lysines are methylated to protect against autolytic degradation. Functionally, trypsin is a serine protease that is specific (cleaves C-terminal of K and R amino acid residues), active in semi-denaturing conditions, and maintains its activity across a selected pH range.^{19,20} The mechanism of trypsin digestion is well understood, and the activity maintains specificity in part due to the highly conserved catalytic triad and the binding pocket arginine. Recent studies have attempted to further refine the rules for trypsin specificity.^{19,21} Regardless, trypsin specificity remains highly conserved across species due to its conserved secondary structure, catalytic motif, and substrate binding pocket.²² This conservation is present in two trypsin sequences of different origin, porcine and bovine, which represent the majority of commercially available MS-grade trypsin enzymes. The combination of biochemical stability and specificity, the suitability for a broad range of sample preparation and analysis methods, and the relatively low cost of the enzyme explain its widespread use in proteomics, although the utility of other enzymes including LysN, LysC, GluC, chymotrypsin, and pepsin have also been explored.^{23,24} There has been an increased interest in developing alternative methods for trypsin digestion;^{20,25,26} however, in-solution and in-gel digestion using a commercially available trypsin remain the primary methods for sample processing in proteomics.

Importantly, protein digestion depends on multiple factors, including the origin of the trypsin, digestion conditions, denaturing conditions, and the presence of post-translational modifications on the protein that may interfere with trypsin digestion.^{27,28} Thus, evaluation of the reproducibility of proteolytic digestion and the degree of digestion completeness is of high importance. In recent reports, variability that is attributed to preanalytical steps such as reduction, alkylation, and trypsin digestion of the target proteins has produced upward of 30% error for intersample variation.²⁶ The referenced study, however, was not designed to determine the specific contribution of protein digestion efficiency toward the measurement accuracy and reproducibility. So far, most recent studies have focused on the development of improved digestion strategies and protocols^{28–37} for complex protein sample analysis. Additionally, recent work has profiled the performance of trypsins from a variety of commercially available sources and for multiple proteins in a mixture.^{38,39} However, there is a need for a more in-depth analysis of trypsin digestion performed under more strictly controlled conditions and using well-defined protein samples. Because MS-based proteomics is poised to become more influential in discovery and diagnostic research for clinical proteomics in the near future, and with protocols for biofluid analysis approaching standardization,^{7,40} it is becoming imperative to develop metrological assays to ascertain performance of preanalytical steps such as protein digestion.

In this work, we comprehensively assayed the performance of six commercially available trypsins of bovine and porcine origin using digests of a highly pure native human serum albumin (HSA). A comprehensive spectral library of tandem mass (MS/MS) spectra of HSA peptides was assembled from thousands of independent analyses of HSA digests and used for robust and comprehensive peptide identification of the HSA digestion products. First, using two of the six trypsins, we investigated the reproducibility of replicate digests and the performance of the LC–MS/MS measurement platform to assess the variability due to sample handling, the instrumentation, and vial-to-vial differences. Second, and as the primary analysis performed in this work, the abundances (or, more precisely, MS signal intensities) of HSA peptides as well as the overall performance metrics, were monitored across HSA digests obtained using all six trypsins and under two different digestion conditions (2 and 18 h digestion time). This analysis revealed significant differences in the propensity of trypsins for missed and irregular cleavages dependent on their origin (bovine or porcine) and digestion time.

■ EXPERIMENTAL METHODS

Trypsins

Six proteomics-grade trypsins were used in the study (M1: G Biosciences no. 786-245, mass spectrometry grade; M2: Princeton Separations no. EN-151, sequencing grade; M3: Promega no. V5111, sequencing grade; M4: G Biosciences no. 786-245B, mass spectrometry grade; M5: Roche Applied Science no. 11418025001, sequencing grade; M6: Worthington Biochemical no. LS02120). These were all TPCK-treated to reduce the chymotryptic activity and methylated for resistance to autolysis. Each of the trypsins was of either porcine (M1–M3) (UniprotKB: P00761) or bovine (M4–M6) (UniprotKB: P00760) origin. Because the main focus of this work was evaluation of the performance of different trypsins for gaining a better understanding of the sources of measurement variability, we

do not disclose a preference for any manufacturer (see Disclaimer).

Human Serum Albumin

HSA purchased from Sigma (catalog number: A3782, purity: > 99%) was used for the initial assessment of reproducibility of replicate digests. The HSA purchased from Lee Biosolutions (catalog number: 101-12, purity: >98%) was used for the main experiment.

Sample Digests

Digests for both the initial reproducibility analysis using two trypsin and the main experiment using all six trypsin were performed under the same conditions, except the digestion time (18 h for the initial reproducibility analysis, 2 and 18 h for the main experiment). In these experiments, 6 mg of HSA (from Sigma in the case of the initial reproducibility analysis and from Lee Biosolutions for the rest of the experiments) was dissolved in 600 μ L of 6 M Urea in 100 mM Tris buffer. Then, 30 μ L of a 200 mM DTT solution was added at room temperature for 1 h to reduce the protein mixture, followed by the addition of 120 μ L of a 200 mM iodoacetamide solution with incubation at room temperature in the dark for 1 h. Then, 120 μ L of a 200 mM DTT solution was added, and samples were incubated for 1 h to eliminate excess iodoacetamine. The resulting protein solution was separated into 145 μ L aliquots. The urea concentration was reduced by diluting the reaction mixture with 755 μ L of water in each vial. Then, a 100 μ L aliquot containing 20 μ g of trypsin from a particular source was added to the vial. Samples were mixed by gentle vortexing, and the digestion was carried out at 37 °C. For each vial, 500 μ L samples were extracted after 2 or 18 h and then quenched with 10 μ L of formic acid (50%) to give a pH <3.

LC–MS/MS

For each run, a 1 μ L aliquot of each 1 μ g/ μ L digest mixtures was injected into a Dionex Ultimate 3000 HPLC (Acclaim pepmap300 column, 150 mm \times 300 μ m, C18, 5 μ m, 300 Å, Dionex, Sunnyvale, CA) and passed through a nanospray source into a Finnigan LTQ mass spectrometer (Thermo Fisher Scientific, Waltham, MA; mass resolution 0.4 m/z). Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. The peptides were eluted by increasing mobile phase B from 1 to 90% over 50 min. Data were collected using a data-dependent mode with a dynamic exclusion time of 20 s. The top eight most abundant precursor ions were selected for ion-trap fragmentation over the m/z range 250–2000. Each sample was analyzed using LC–MS/MS in triplicate.

HSA Spectral Library

The reference spectral library (Dong et al., in preparation) was compiled from over 3000 digest runs for a wide range of digestion conditions. A set of peptide identification search engines was used to find all identifiable peptide products. This included multiple missed cleavages and up to six charge states as well as semitryptic products (from in-source fragmentation as well as those from irregular cleavages). A wide range of common modifications were included in the search, ranging from sodiated aspartic acid residues to cyclization of N-terminal cysteine. Additional, less common modifications were found using programs capable of finding untargeted modifications (i.e., so-called ‘blind’ mode). High-scoring identifications (passing a 5% local FDR filter) from all analyses were then combined, and their MS/MS spectra were clustered to produce

consensus spectra.^{40,41} These consensus spectra were then subjected to a battery of quality assessment filters. These, for example, eliminated MS/MS spectra corresponding to MS1 features that were rarely detected or identifications with having multiple rare modifications, inconsistent retention time, or with large mass errors. Summary statistics for the HSA library are shown in Supplementary Figure 1 in the Supporting Information.

Data Analysis

Raw data files were converted to mzXML and MGF using the ReadW4Mascot2 converter^{9,41} (an extension of ReadW.exe). MSPepSearch (v.0.9)⁹ was used to match spectra against the HSA library described above. Search tolerances were 1.8 m/z units for the precursor and 0.8 m/z units for fragment peaks. A minimum MSPepSearch score of 450 was required for inclusion of identified spectra in the subsequent analysis using `nistms_metrics.exe`, `ReadW4Mascot2.exe`, `ProMS.exe`, `MSPepSearch.exe`, `MergePepResults.exe`, and `NISTMS_metrics.exe` were all run, and the data were combined using the NISTMSQC v1.2.0 data analysis pipeline.⁹ These tools are available for download at <http://peptide.nist.gov>. Quality metrics and their deviations for each raw data file were also calculated. Intraserie and interserie deviations were reported as coefficients of variation (CV). CVs for the LC–MS/MS replicates were computed using the values across three LC–MS/MS replicates. The interserie (the comparison between all the trypsin) and intraserie (the comparison only between trypsin of the same origin, porcine or bovine) CVs were computed using the mean values from each sample. All data were uploaded into a custom MySQL (v.5.0.95) database, and additional statistics were calculated using PHP (v.5.3.3) scripts. Any further analysis and plots were completed using R (2.14) or Microsoft Excel 2007.

Peptide intensities (area calculations from extracted ion chromatograms) were calculated using ProMS (v.0.9). Peptide intensities were normalized to the total intensity of all ions identified in each LC–MS/MS run. The intensity for each unique peptide sequence in each LC–MS/MS run was calculated as the sum of the normalized intensities of all identified peptide ions detected for that sequence in that run (i.e., summing the normalized intensities of peptide ions containing different modifications and or identified in different charge states). Finally, the intensity of each unique peptide sequence in each trypsin digest (and each digestion time point) was computed as the average intensity across the three LC–MS/MS replicates for each sample. In doing so, missing values were dropped from the computation, with the mean computed only from observed values. As such, the reported CVs were included only for ions with at least two measurements (out of three LC–MS/MS replicates) per sample. Comprehensive lists of the peptides identified in the main experiment, including their abundances, are included in Supplementary Tables 1 and 2 in the Supporting Information for the 2 and 18 h digests, respectively. The lists of all the ions detected in the analysis for each of the six trypsin and used in computations of peptide intensities are included in Supplementary Tables 3 and 4 in the Supporting Information.

RESULTS AND DISCUSSION

Overview

An overview of the analysis is shown in Figure 1. The main application of the data analysis pipeline and visualization tools

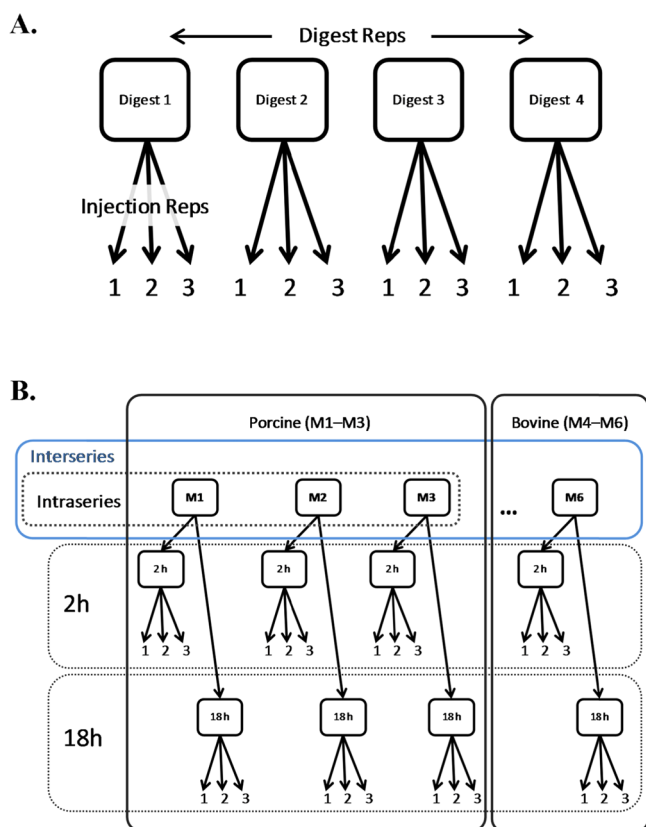


Figure 1. Overview of the experiments. (A) Digest replicates were performed on three separate days, four times, and analyzed by LC-MS/MS in triplicate. (B) Trypsin digests were performed using six different trypsin types of bovine or porcine origin, for 2 or 18 h digestion time, and then analyzed by LC-MS/MS in triplicate. Interseries analysis describes the analysis of trypsin digests across all six different trypsin types; intraseries analysis is done separately across three porcine (M1–M3) or three bovine (M4–M6) trypsin digests.

assembled in this work was to evaluate the performance of trypsin from different manufacturers (M1–M3: porcine, M4–M6: bovine; see the Methods for details). As such, digests of a single substrate, HSA, were used to test individual enzyme performance. This ensured the greatest sensitivity to the differences between the trypsin types. The initial analysis (Figure 1A) investigated the reproducibility of replicate digests performed on different days. For this part, two of the six trypsin types (M2 and M5) were used to digest HSA for 18 h four times on different days and analyzed in triplicate using LC-MS/MS. For the primary experiment (Figure 1B), single digests for each of the six trypsin types were used, with 2 and 18 h digestion time, and analyzed using LC-MS/MS in triplicate. The resulting data were used for the following three primary analyses: (1) analysis across all trypsin types regardless of origin (interseries M1–M6), (2) separate analysis for the two trypsin sequences (intraseries, porcine: M1–M3 and bovine: M4–M6), and (3) comparison between the two digestion times (2 vs 18 h digestion).

Digestion Reproducibility and LC-MS/MS Platform Stability

To assess the reproducibility of the digests and the stability of the entire measurement platform (Figure 1A), we analyzed repeated digest using the M2 and M5 trypsin types by LC-MS/MS in triplicate. The acquired data were processed using the NIST

MSQC tools (see Methods). The results were summarized using various metrics as described in Rudnick et al.⁹ The metrics of most relevance to the aims of this study are listed in Tables 1 and 2. These included the total number of MS/MS scans acquired (DS2-B), the number of identified unique ions (P-2B) and peptides (P-2C), the ratios of the number of ions observed in different charge states compared with that in charge state $z = 2^+$ (IS-3A through IS-3C), and the mean length of peptides (PL-1 to PL-4) detected for each charge state (Table 1).

For the LC-MS/MS replicates, the CVs for the DS2-B metric were consistently <2.1%, and the CVs for the total detected unique ions and peptides never exceeded 3.9% CV. Higher CVs (up to 14.1%) were reported for the IS-3 metrics. This reflects the more variable nature of the ionization process with regard to higher charge ions ($z = 3^+$ or higher). However, it is $z = 2^+$ ions that account for the majority (~70%) of the total detected ions. The PL-1 to PL-4 CVs were below 2.7%. The CVs for the repeated digests completed on separate days, for each of the two trypsin types, were similar or increased slightly compared with the CVs for the LC-MS/MS replicates of the same digest. For the digest replicates, the DS2-B CVs were 2.4 and 2.1% for the M2 and M5 trypsin types, respectively. Excluding the IS-3 metrics, the CVs increased from an average of 1.3% for the LC-MS/MS replicates to an average of 2.2% for digest replicates. Taken together, the data showed the high reproducibility of the LC-MS/MS measurement platform and that performing replicate digests under the same conditions was not a major source of variability in this study. Thus, the subsequent analysis across all six trypsin types, described below, was performed using a single digest for each trypsin.

Six Trypsins Study

The main experiment of this study, performed using six different trypsin types (Figure 1B), is discussed in the remainder of this manuscript. Reported measures of variability (Table 2) were low for most metrics when considering the LC-MS/MS replicates (labeled M1–M6 CV columns in Table 2), as expected based on the results presented above. The greatest variances were observed in the analysis across all trypsin types (labeled as “interseries” in Table 2). The total number of MS/MS scans remained stable throughout the experiment for all comparisons (2 h vs 18 h, bovine vs porcine, interseries). The numbers of unique peptide sequences (P-2C) and unique peptide ions (P-2B) that were produced at 2 h were 121.4 ± 17.2 and 312.9 ± 10.4 , respectively, and decreased slightly at 18 h. As previously discussed, the highest CVs were observed for the IS-3A-C metrics at both time points.

On the basis of principal components analysis (PCA), discussed later in this manuscript, the origin of the trypsin (bovine vs porcine) was found to be a significant factor contributing to the overall variability of peptide abundances. At 2 h, the classification of the trypsin types into two subclasses resulted in significantly decreased CVs within each class (intraseries analysis in Figure 1B; see Table 2, columns ‘porcine’ and ‘bovine’) as compared with the interseries analysis. For example, the CVs were <7% for the reported number of unique peptides when analyzed separately for each subclass (compared with 17.2% for the interseries analysis). Interestingly, the mean values of the total unique peptides and total unique ions were 33.1 and 14.7% higher for the bovine trypsin group than for the porcine group (Table 2, column ‘Bov/Por’). The bovine trypsin types also produced more peptides that ionized into high charge state ions (3^+ and 4^+) and with lower CVs (IS-3B and IS-3C metrics). At 18 h, significant differences were

Table 1. Metrics for Analysis of Reproducibility: Repeated Digest Experiment^a

metric	description	M2 digest 1 mean CV	M2 digest 2 mean CV	M2 digest 3 mean CV	M2 digest 4 mean CV	M5 digest 1 mean CV	M5 digest 2 mean CV	M5 digest 3 mean CV	M5 digest 4 mean CV	M2 mean CV	M5 mean CV
DS2-B	MS/MS scans	10259.7	2.1	10786.7	0.2	10796.0	0.5	10497.3	0.3	10373.3	0.7
P-2C	unique peptides	109.7	2.1	121.0	0.8	119.3	1.7	116.7	1.3	127.7	1.8
P-2B	unique ions	343.7	3.7	350.7	1.4	352.7	0.9	342.7	2.1	394.3	3.9
IS-3A	ratio 1z/2z	0.3	5.4	0.3	12.5	0.3	8.4	0.3	2.1	0.3	8.5
IS-3B	ratio 3z/2z	0.7	6.5	0.8	5.8	0.9	5.2	0.8	7.9	0.9	8.9
IS-3C	ratio 4z/2z	0.4	3.1	0.4	2.7	0.5	5.7	0.4	0.5	0.5	3.8
PL-1	AVG length z = 1	7.8	1.0	7.7	3.1	7.5	2.3	7.5	0.9	7.5	3.1
PL-2	AVG length z = 2	10.1	0.6	9.9	0.7	9.9	1.2	9.9	1.3	10.3	1.0
PL-3	AVG length z = 3	15.6	1.8	14.9	1.1	14.6	0.4	15.0	1.2	15.4	1.9
PL-4	AVG length z = 4	22.0	0.7	21.3	1.0	21.1	2.4	22.1	1.3	22.0	1.4

^aMetrics from Rudnick et al.⁹ used in this work. Mean and % CV of each metric are reported for either the M2 or M5 trypsin, four digests for each trypsin. Mean and % CVs were calculated for each replicate digest (labeled digest 1 through digest 4). Additionally, the mean and % CVs were calculated for each trypsin using the mean values for each replicate digest (far right two columns). DS: dynamic sampling, P: peptide, IS: ion source, PL: peptide length.

observed for the number of unique ions. These increased, compared with 2 h digestion time, by ~20 and 26% for the porcine and bovine trypsins, respectively, and the % difference (bovine vs porcine) increased by 6%. The average peptide lengths were slightly different between bovine and porcine trypsin digests: bovine trypsins produced longer peptides when considering the identifications resulting from high-charge-state peptide ions (e.g., longer by 2.5 amino acids, PL-4 metric).

The increase in the average peptide length and in the number of peptide identifications from high charge state ions in bovine versus porcine digests suggested that bovine trypsins were more likely to produce peptides containing missed cleavages. Protease specificity is partially determined by its substrate binding pocket, the surface fit between the substrate and enzyme, and the primary sequence of the substrate.^{19,42} Thus, the two enzyme sequences (porcine vs bovine) may exhibit differences in their substrate affinities that should be observable in these data. To test this further, different trypsins were compared with respect to the types of cleavage that led to the formation of the peptides (Figure 2). The total numbers of identified unique peptide sequences were counted considering whether they were fully tryptic peptides with no missed cleavages (FT), fully tryptic peptides containing one or more missed cleavage (MC), or semitryptic peptides (ST).

At 2 h, the interseries analysis (across all trypsins, M1–M6, Figure 2A) demonstrated that total counts of FT peptides were consistent between all trypsins ($43.8 \pm 0.2\%$ unique peptide identifications). Peptides classified as MC and ST were significantly more varied: 84.2 ± 21.0 and $59.9 \pm 12.0\%$, respectively. Most notable were a higher number of MC peptides in the M4–M6 digests and more ST peptides in the M2 digest. At 18 h (Figure 2B), the number of MC peptides decreased slightly (14 fewer peptides) for all trypsins, but the differences between the M1–M3 and M4–M6 trypsins became more visible. The number of ST peptides increased in 18 h digests compared with 2 h digests for all trypsins, and the M2 trypsin produced a more marked increase in the number of these peptides (~50% more ST peptides over the other trypsins).

Grouping the trypsins by their origin (porcine or bovine) made these trends even more apparent (Figure 2C). The bovine group collectively produced a significantly higher number of unique MC peptides ($66.2 \pm 9.9\%$ in porcine vs $102.1 \pm 9.4\%$ in bovine; 54% more in bovine), whereas the porcine trypsins produced slightly more ST peptides ($62.1 \pm 27.8\%$ in porcine vs $57.0 \pm 10.5\%$ in bovine; 9% more in porcine). The number of FT peptides stayed the same across trypsins and for the 18 versus 2 h digests. The total numbers of MC peptides were higher for the bovine versus porcine trypsins regardless of the digestion time but decreased with increased digestion time for both classes of trypsin (down 33% in porcine and 5.6% in bovine). ST peptides increased 62% in porcine and 29% in the bovine trypsin digests from 2 to 18 h.

Analysis of Peptide Abundances

The above analysis of the counts of the identified peptides was extended by considering the quantitative measure of peptide intensity. Here the intensities of peptides were estimated based on the intensities of peptide ions extracted from the MS1 data using the ProMS tool of the NIST MSQC pipeline, which were then summed for each unique peptide sequence. The intensities were normalized to the total intensity of all peptides in the LC–MS/MS run. (See the Methods for details.) As such, the peptide intensity numbers discussed below are given as the fraction of the total HSA intensity detected in each sample.

Table 2. Metrics for Analysis of Reproducibility: Six Trypsins Study^a

(A) 2 h																				
metric	description	M1		M2		M3		M4		M5		M6		interseries		porcine		bovine		Bov/ Por % diff
		mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	
DS2-B	MS/MS scans	10081.7	0.7	10085.3	0.5	10073.3	0.6	10187.0	0.8	10099.0	0.5	10121.3	0.5	10107.9	0.4	10080.1	0.1	10135.8	0.5	0.6
P-2C	unique peptides	108.7	3.0	107.7	4.2	96.0	1.8	139.3	2.5	130.3	1.9	146.3	2.1	121.4	17.2	104.1	6.8	138.6	5.8	33.1
P-2B	unique ions	312.0	2.0	296.0	3.3	266.3	1.5	333.3	2.9	320.0	1.7	349.7	0.9	312.9	10.4	291.4	8.0	334.3	4.4	14.7
IS-3A	ratio 1z/2z	0.5	5.2	0.5	3.2	0.4	3.4	0.3	2.4	0.4	6.9	0.4	6.2	0.4	11.0	0.4	9.2	0.4	8.3	-14.2
IS-3B	ratio 3z/2z	1.0	0.5	1.0	10.7	0.8	4.1	1.2	2.1	1.1	6.5	1.1	6.0	1.0	15.1	0.9	13.9	1.1	4.1	21.6
IS-3C	ratio 4z/2z	0.5	8.3	0.6	3.4	0.4	9.1	0.7	4.0	0.7	6.6	0.7	10.8	0.6	22.1	0.5	15.7	0.7	1.9	42.9
PL-1	AVG length z = 1	7.9	0.9	7.7	1.9	7.8	1.7	7.5	1.3	7.6	1.6	7.5	0.8	7.7	1.6	7.8	0.8	7.5	0.6	-3.3
PL-2	AVG length z = 2	10.8	1.2	11.0	2.0	10.7	1.5	10.8	1.2	10.9	0.5	10.9	1.4	10.9	1.0	10.8	1.3	10.9	0.5	0.5
PL-3	AVG length z = 3	16.1	0.7	16.1	2.0	15.8	3.0	16.1	1.2	16.1	3.4	16.6	1.2	16.2	1.8	16.0	1.0	16.3	1.8	1.7
PL-4	AVG length z = 4	22.0	0.6	21.7	0.5	22.1	0.3	22.3	1.7	22.3	2.1	22.1	1.5	22.1	1.1	21.9	0.9	22.2	0.4	1.5

(B) 18 h																				
metric	description	M1		M2		M3		M4		M5		M6		interseries		porcine		bovine		Bov/ Por % diff
		mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	mean	CV	
DS2-B	MS/MS scans	10026.0	0.9	10115.0	1.0	10077.7	1.2	10085.7	1.2	10143.0	0.7	10113.3	1.4	10093.5	0.3	10072.9	0.4	10114.0	0.3	0.4
P-2C	unique peptides	98.3	2.6	90.7	4.2	87.7	1.7	142.7	1.6	132.0	0.8	148.7	0.4	116.7	24.2	92.2	5.9	141.1	6.0	53.0
P-2B	unique ions	370.3	1.6	357.3	1.3	320.0	1.9	432.7	1.8	406.7	3.1	427.7	3.2	385.8	12.5	349.2	7.5	422.4	3.3	21.0
IS-3A	ratio 1z/2z	0.4	4.9	0.4	0.5	0.4	3.9	0.3	2.9	0.4	1.5	0.3	7.1	0.4	6.5	0.4	5.7	0.3	4.5	-8.3
IS-3B	ratio 3z/2z	0.8	0.7	0.8	4.9	0.7	3.9	1.2	2.8	1.0	3.7	1.1	3.3	0.9	22.7	0.8	12.5	1.1	6.7	43.7
IS-3C	ratio 4z/2z	0.3	4.9	0.4	3.4	0.3	4.3	0.7	6.4	0.5	3.2	0.7	6.5	0.5	35.9	0.3	13.0	0.6	11.7	91.7
PL-1	AVG length z = 1	7.7	0.8	7.8	1.2	8.0	0.5	7.5	1.4	7.6	1.5	7.6	0.5	7.7	2.6	7.8	2.0	7.6	0.7	-3.4
PL-2	AVG length z = 2	10.8	0.8	10.8	1.5	10.6	1.2	11.1	0.3	11.1	0.6	11.2	1.8	10.9	2.2	10.7	0.9	11.1	0.4	3.6
PL-3	AVG length z = 3	15.9	1.0	15.7	0.9	15.8	0.9	16.4	0.8	16.5	1.1	16.7	1.0	16.2	2.7	15.8	0.6	16.5	0.9	4.5
PL-4	AVG length z = 4	21.1	0.2	20.4	0.9	20.4	0.3	23.0	1.3	23.2	0.9	23.3	0.8	21.9	6.8	20.6	1.8	23.2	0.6	12.2

^aMean and % CV of each metric are reported for each trypsin (M1–M6). Interseries mean \pm % CV are calculated for all trypsin combined. Porcine or bovine \pm % CV are the reported values for the intraseries analysis (between porcine (M1–M3) or bovine (M4–M6) trypsin). (A) 2 h digests. (B) 18 h digests. Bov/Por % change is the percent difference between the mean reported values for the bovine and porcine digests. DS: dynamic sampling; P: peptide; IS: ion source; PL: peptide length.

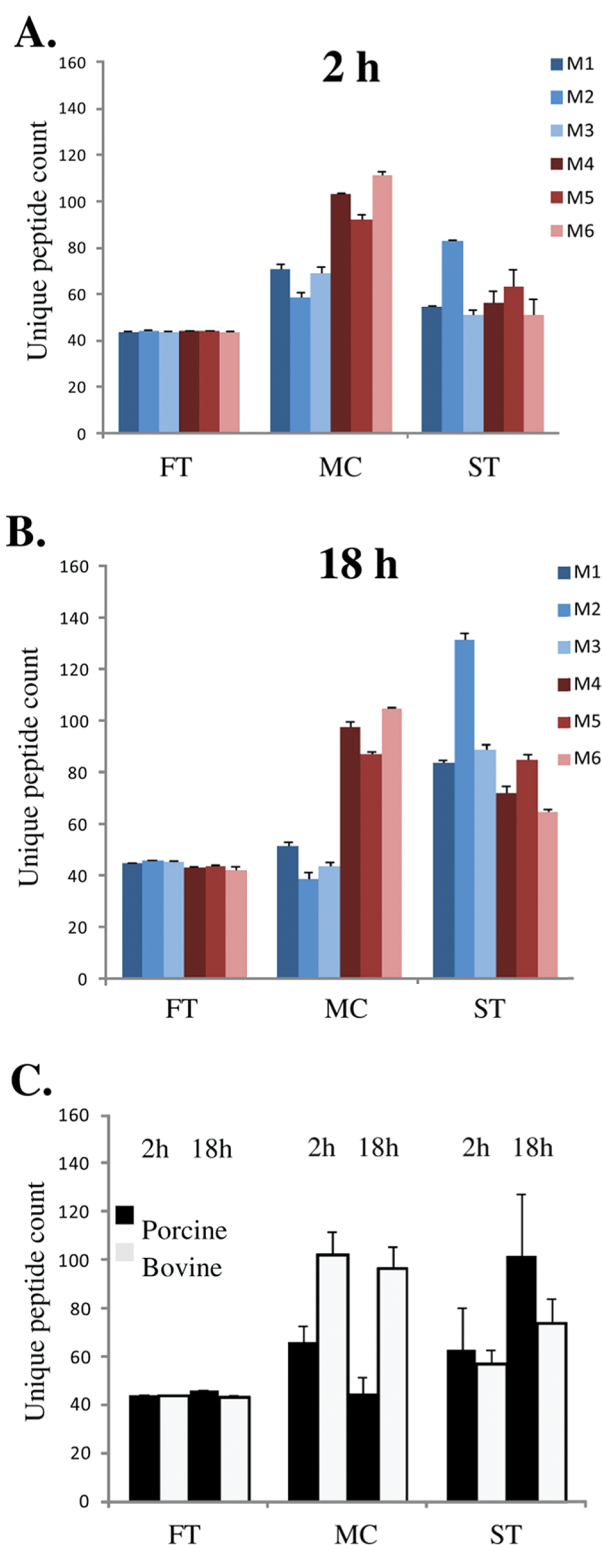


Figure 2. Unique peptide identifications. Unique peptide counts (mean \pm standard deviation across three replicates) in three different peptide categories (FT: fully tryptic, MC: missed cleavage, ST: semitryptic) are plotted for each trypsin (M1–M6) for both the 2 h (A) and 18 h (B) digestion time points. (C) Same as above, data grouped by porcine or bovine trypsin category (mean \pm standard deviation across all trypsins and replicates for each trypsin category), 2 and 18 h digests. M1–M3: porcine trypsins, M4–M6: bovine trypsins.

At 2 h, the peptide intensity trends (Figure 3A) were similar to those observed for the unique peptide counts (Figure 2A), except that the differences between FT peptide intensities in bovine and porcine trypsins became discernible (M1–M3: $0.34 \pm 1.4\%$; M4–M6: $0.25 \pm 8.0\%$). Relative to the 2 h digests, FT peptide intensities in 18 h digests (Figure 3B) increased by 15.3 to $0.34 \pm 20.6\%$, MC intensities decreased 17.1 to $0.44 \pm 30.0\%$, and ST peptide intensities increased by 28.5 to $0.21 \pm 33.3\%$. The MC and ST peptides were more variably produced than FT peptides across all the trypsins. Similar to the count data shown in Figure 2, there were significantly different intensities of MC and ST peptides between the M1–M3 and M4–M6 trypsins. These differences become more apparent in Figure 3C. In porcine trypsin digests, the total FT and ST peptide intensity increased with longer digestion time by 19.4 and 41.3%, respectively. This was significantly more than the corresponding 9.7 and 12.1% increases in bovine trypsin digests. At the same time, the intensity of MC peptides decreased greatly in porcine (29.9%) but only slightly (6.7%) in bovine trypsin digests. The differences between the bovine and porcine trypsins generally were more pronounced at 18 h.

Considering all of the data together, FT peptides as a category were found to be the most reproducible type of peptides. The increase in the intensity of FT peptides in 18 versus 2 h digests, accompanied by a corresponding decrease in the number and intensities of MC peptides, indicated more complete HSA digestion with longer digestion time. The increase in both the intensity and in the number of unique ST peptides with increased digestion time may indicate an increased probability of trypsin producing an irregular cleavage (non K/R). However, this could also be attributed to other factors such as peptide degradation after trypsin digestion or increased activity of contaminating enzymes such as chymotrypsin. The ST peptides significantly contributed to the total summed HSA intensity ($\sim 21\%$), and the number of identified unique ST peptides was higher than that of MC or FT peptides. It should also be noted that the high number of identified ST peptides in these data was due to low sample complexity and is not representative of the numbers of ST peptides observed in a typical analysis of complex protein samples.^{38,43} A detailed list of the detected peptides and their intensities for both the 2 and 18 h digests is included in Supplementary Tables 1 and 2 in the Supporting Information.

■ ANALYSIS OF TRYPSIN DIGESTION USING PRINCIPAL COMPONENTS ANALYSIS

We sought to use a statistical and visual approach to highlight the global peptide intensity trends. To this end, we first performed a PCA (Figure 4). The input in the exploratory PCA analysis was a matrix consisting of 267 peptide intensity measurements (peptides identified and quantified in both time points). The PCA analysis identified three statistically significant principal components (PCs). The first and most significant component (PC1; explained 89% of the total peptide intensity variance) summarized the differences between the intensities of different peptides in these data. (See PC1 vector coefficients listed in Supplementary Figure 2 in the Supporting Information). This component could be explained by the biases in the entire measurement system, including the effect of the physiochemical properties of a peptide on its ionization efficiency. The second component, PC2, explained 7% of the variance after considering the variance explained by PC1. This component accounted for the differences between the porcine and bovine trypsins. (See PC2 vector coefficients listed in Supplementary Figure 2 in the Supporting Information).

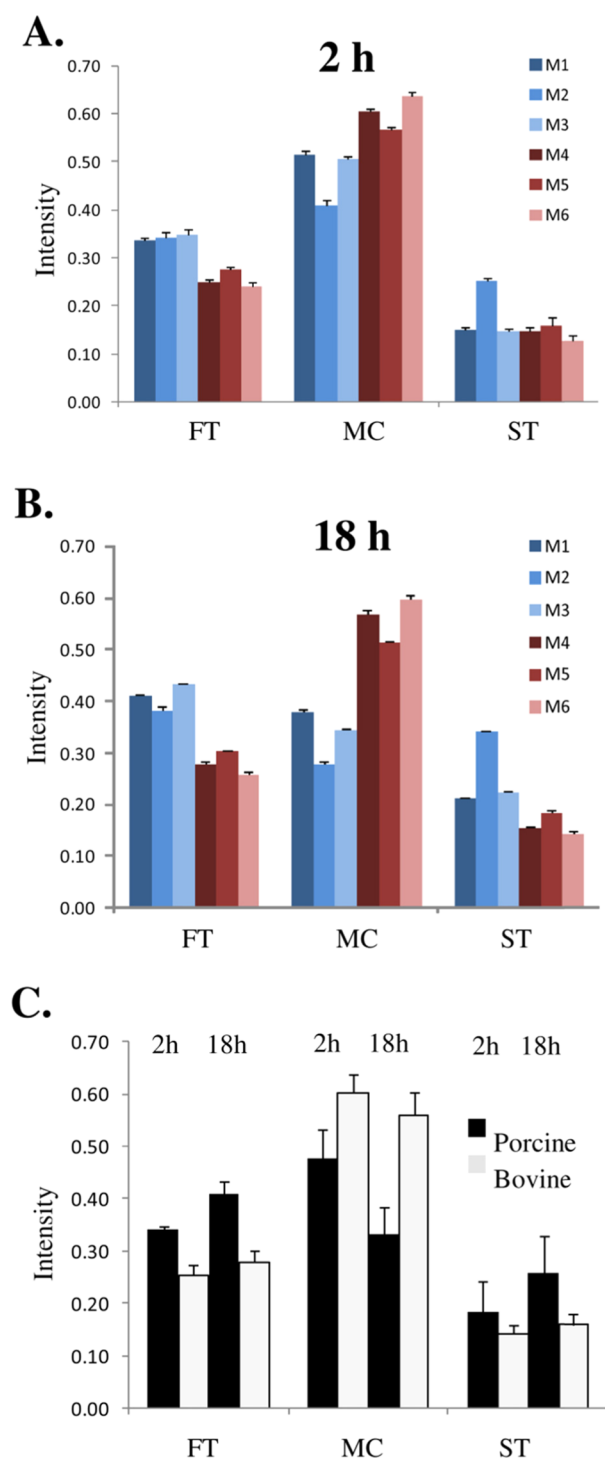


Figure 3. Peptide intensities. Peptide intensities computed using normalized peptide ion intensities (mean \pm standard deviation across three LC-MS/MS analysis) in three different peptide categories (FT: fully tryptic, MC: missed cleavage, ST: semitryptic) are plotted for each trypsin (M1–M6) for both the 2 (A) and 18 h (B) digestion time points. Intensities are shown as the fraction of the total intensity of all HSA peptides (for each digest) contributed by peptides from a particular category. (C) Same as above, data grouped by porcine or bovine trypsin category (mean \pm standard deviation across all trypsins and replicates for each trypsin category), 2 and 18 h digests. M1–M3: porcine trypsins, M4–M6: bovine trypsins.

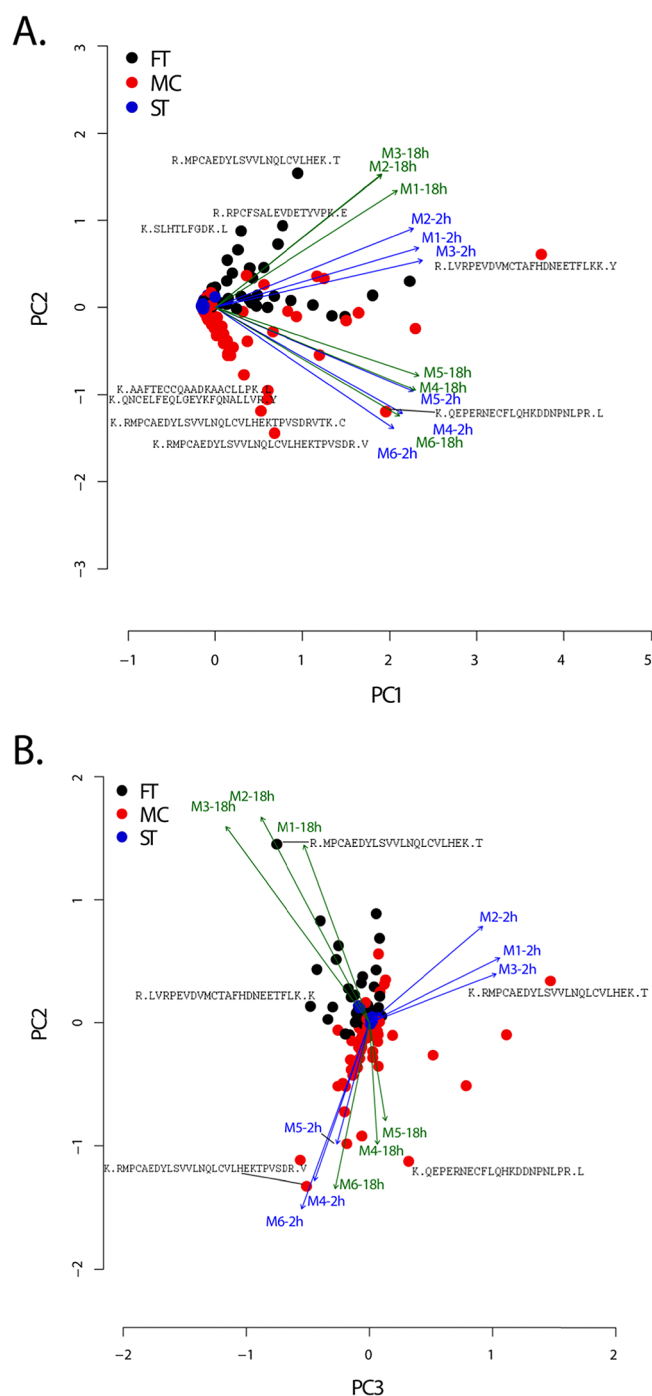


Figure 4. Principle components analysis. (A) PC1 versus PC2 plot. PC2 component describes the differences between peptide intensities in digests using porcine (M1–M3) versus bovine (M4–M6) trypsins. Selected peptides whose intensities contribute most significantly toward differentiating between the bovine and porcine trypsins are indicated. (B) Same as previous, except PC3 versus PC2 are plotted.

Figure 4A plots the data using the first two principal components and visually demonstrates the separation between bovine and porcine trypsins for both 2 and 18 h digests. While some of the MC peptides were produced with greater variance and were low in intensity (centered around zero on PC1 and PC2 axes), several MC peptides were very abundant (high PC1 values) and were reliably detected by most or all trypsins (low PC2 values). This was exemplified by the MC peptide RLVRPEVDVMCTAFHDNEETFLKK.Y, the most abundant

Table 3. Peptides with Most Altered Intensities between Bovine and Porcine Trypsin Digests^a

2 h														18 h													
peptide	flanking residues	Cat	Pos	M1	M2	M3	M4	M5	M6	μ por	μ bov	$FC_{bov/por}$	peptide	flanking residues	Cat	Pos	M1	M2	M3	M4	M5	M6	μ por	μ bov	$FC_{bov/por}$		
RMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK	KC	MC	468	0.10	0.10		31.68	28.33	24.34	0.10	28.12	89.7	RMPCAEDYLSVVLNQLCVLHEKTPVSDR	KE	MC	468	0.19			18.90	31.06	34.58	0.19	28.18	91.1		
LDELDEGKASSAK	KQ	MC	205			0.18	1 1.29	6.99	10.77	0.18	9.68	31.9	TCVADESAENCDSLHTLFGDKLCTVATLR	KF	MC	75				14.40	11.37	16.85		14.21	57.8		
RMPCAEDYLSVVLNQLCVLHEKTPVSDR	KV	MC	468	0.92	0.28	0.68	18.93	23.65	26.49	0.62	23.03	26.6	RMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK	KC	MC	468				17.78	13.86	10.08		13.91	56.6		
AEFAESKLVTDLTK	KV	MC	249				5.83	2.68	9.63		6.05	25.2	QNCLEFEQGEYKFQNALVLR	KY	MC	413	0.21	0.18	0.09	19.96	15.45	23.17	0.16	19.52	48.4		
VFDEFKPLVEEPQNLKQNCLEFEQGEYKFQNALVLR	KY	ML	396	0.23		0.27	11.12	4.10	14.32	0.25	9.85	24.2	DVCKNTAEAKDVFELGMFLYEYARR	KL	MC	557	0.11	0.07	0.12	13.11	19.29	17.65	0.10	16.68	48.2		
AFKAWAVAR	RL	MC	233				4.48	0.10	5.07		3.22	13.9	AAFTCCQAADKAACLLPK	KL	MC	186	0.28	0.18	0.15	18.94	21.11	17.51	0.20	19.19	42.8		
MPCAEDYLSVVLNQLCVLHEKTPVSDR	RV	MC	469	0.50	0.22	0.30	7.03	5.72	8.20	0.34	6.98	12.3	LDELDEGKASSAK	KQ	MC	205				10.03	6.07	10.46		8.85	36.4		
LSQRFPAEFAEVS	RL	MC	242				2.59	0.34	4.86		2.60	11.4	VFDEFKPLVEEPQNLKQNCLEFEQGEYKFQNALVLR	KY	MC	596				7.24	3.71	14.49		8.48	34.9		
LVPRVDVMCTAFHDNEETFLKKVLYEIR	RR	MC	138	0.14		0.11	3.49	1.46	5.63	0.13	3.53	11.3	YKAAFTCCQAADKAACLLPK	RL	MC	184				6.80	2.74	6.58		5.37	22.5		
MPCAEDYLSVVLNQLCVLHEKTPVSDRVTK	RC	MC	469	0.06			3.07	1.94	3.35	0.06	2.78	11.2															
VFDEFKPLVEEPQNLKQNCLEFEQGEYK	KF	MC	596	0.20	0.07	0.21	4.71	3.43	3.81	0.16	3.98	10.3															
CKKADDKETCFAEEGKK	KL	MC	581	1.50	0.05	1.17	9.35	9.81	13.47	0.91	10.87	9.6															
AACLLPKLDELDEGKASSAK	KQ	MC	198				1.50	0.74	2.82		1.69	7.8															
LVTDLTKVHTECHGDLLECADDRADLAK	KY	MC	257				1.02	1.04	2.64		1.57	7.3															
YLYEIARR	KH	MC	161	0.66		0.55	4.99	3.42	4.59	0.61	4.33	7.0															
DDNPNLPR	KL	FT	130	1.74	4.54	1.84	0.64	1.80	0.79	2.71	1.08	0.45															
TCVADESAENC DK	KS	FT	75	3.51	3.48	2.17	1.14	1.41	0.88	3.06	1.14	0.42															
FQNALVLR	KY	FT	426	19.23	24.30	19.07	8.08	12.26	4.71	20.87	8.35	0.41															
LDELDEGK	KA	MC	205	12.42	13.47	11.52	4.19	7.61	2.96	12.47	4.92	0.41															
AACLLPK	KL	FT	198	12.24	15.97	11.43	5.10	6.18	4.07	13.21	5.12	0.40															
QNCLEFEQGEYK	KF	FT	415	7.19	8.84	7.26	2.36	4.19	1.35	7.76	2.64	0.36															
NECFLOHKDDNPNLPR	RL	MC	122	1.90	3.13	1.92	0.69	0.62	0.40	2.32	0.57	0.32															
RPCFSALEVDVETVYPK	RE	FT	508	21.83	23.15	21.98	4.37	12.34	3.89	22.32	6.87	0.32															
AAFTCCQAADK	KA	FT	186	6.44	9.46	5.87	1.85	2.54	1.72	7.26	2.04	0.30															
RMPCAEDYLSVVLNQLCVLHEK	KI	MC	468	39.40	44.60	41.90	12.92	14.25	10.06	41.97	12.41	0.30															
LCTVATLR	KE	FT	97	7.02	11.80	6.64	1.57	3.63	0.89	8.49	2.03	0.26															
MPCAEDYLSVVLNQLCVLHEK	RT	FT	469	19.29	19.70	22.07	5.32	5.46	4.30	20.35	5.03	0.26															
EFNAETFTFHADICTLSEK	KE	FT	524	4.75	4.99	6.10	0.50	0.86	0.33	5.28	0.56	0.15															
SLHTLFGDK	KL	FT	88	6.81	12.45	6.77	0.51	1.70	0.34	8.68	0.85	0.12															
TPVSDR	KV	FT	490	4.02	4.82	3.96	0.71			4.27	0.71	0.11															

Table 3. continued

peptide	flanking residues	Cat	Pos	18 h							μ por	μ bov	FC _{bov/por}
				M1	M2	M3	M4	M5	M6				
CKKADDKETCFAEEGKK	K.L	MC	581				7.23	8.11	0.40			8.96	22.0
AEFAEVSKLVTDLTK	K.V	MC	249				3.85	1.90	8.37			4.70	19.8
MPCAEEDYLSVNLQCVLHEKTPVSDR	R.V	MC	469	0.44		0.12	6.10	7.23	11.03		0.28	8.12	19.1
CCTESLVNRRPCFSALEVDETYVPK	K.E	MC	499				6.49	0.85	6.24			4.53	19.1
YLVEIARR	K.H	MC	161				4.22	2.92	4.66			3.92	16.8
TPVSDRVTK	K.C	MC	490	0.08			4.62	4.08	3.55		0.08	4.08	15.7
RPCFSALEVDETYVPK	RE	FT	508	28.07	29.41	27.78	6.16	13.53	4.71	28.42		8.13	0.29
QNCLEFEQLGEYK	K.F	FT	413	10.75	12.36	12.18	2.67	4.35	1.44	11.77		2.82	0.26
ALVLIAF	K.A	ST	44	0.65	1.17	0.57				0.80			0.24
QEPERNECFLOHK	K.D	MC	117	2.51	4.31	2.82	0.25	0.66	0.35	3.21		0.42	0.19
DDNPNLPR	K.L	FT	130	4.64	8.97	6.23	0.70	1.45	0.78	6.61		0.97	0.18
MPCAEEDYLSVNLQCVLHEK	R.T	FT	469	44.47	41.92	52.23	9.89	7.98	6.25	46.21		8.04	0.18
NECFLOHKDDNPNLPR	R.L	MC	122	4.84	3.45	4.78	0.58	0.63	0.42	4.36		0.54	0.17
VHTECCHGDLLECADDR	K.A	FT	264	4.26	7.76	6.40	0.83	1.03	0.49	6.14		0.78	0.16
NECFLOHK	R.D	FT	122	0.62	2.54	1.00				1.38			0.15
FSALEVDETYVPK	C.E	ST	511	2.51	4.72	5.90	0.41	0.37	0.28	3.71		0.35	0.15
LCTVATLR	K.E	FT	97	15.36	20.28	17.91	1.68	3.28	0.87	17.85		1.94	0.12
TFHADICTLSEK	F.E	ST	531	0.07	5.32	0.12				1.84			0.12
EFNAETFTFHADICTLSEK	K.E	FT	524	16.57	6.43	20.79	1.13	1.57	0.71	14.59		1.14	0.09
SLHTLFGDK	K.L	FT	88	20.39	25.28	22.26	1.31	2.11	0.25	22.65		1.23	0.06
TPVSDR	K.V	FT	490	4.63	4.54	4.01				4.39			0.05

^aPeptide intensities are reported for each trypsin (M1–M3: porcine, M4–M6: bovine), along with mean intensity for the porcine (μ_{por}) and bovine (μ_{bov}) group. Position is the start position for each peptide in the sequence of HSA. Flanking residues are the immediate N or C terminal residues of the peptide in the protein sequence. $FC_{bov/por}$ indicates fold change for the mean intensity in the bovine versus porcine digests. To account for missing values, fold changes were computed after the addition of a 0.25 background factor to all M1–M6 intensities. Top 30 peptides are listed for each time point (the 15 peptides with the highest $FC_{bov/por}$ and no missing values across bovine trypsin digests and the 15 peptides with the lowest $FC_{bov/por}$ and no missing values across porcine trypsin digests). (A) 2 h digests. (B) 18 h digests. FT: fully tryptic, MC: missed cleavage, ST: semi-tryptic.

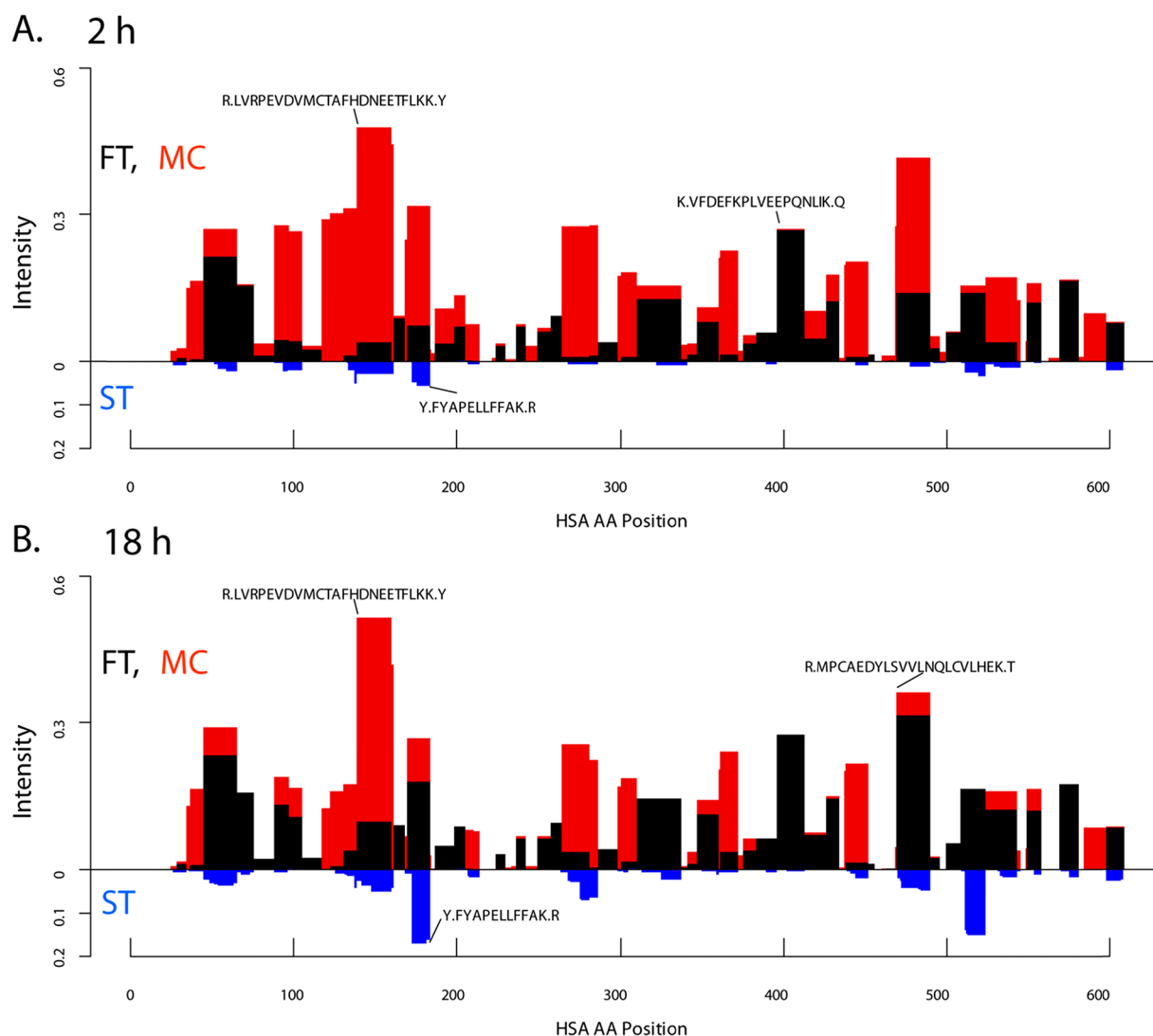


Figure 5. Peptide intensity as a function of amino acid position along the HSA protein sequence. Stacked bars (y axis) represent the summed peptide intensities of fully tryptic (FT, black), missed cleavages (MC, red), and semitryptic (ST, blue) peptides covering a particular amino acid position (x axis). ST peptides are shown on the separate scale from FT and MC peptides. Data for M3 trypsin only, 2 h (A) and 18 h (B) digestion time points. Selected peptide sequences with the highest measured intensity for each of the cleavage rules are indicated and discussed in the text.

peptide in the data. Both FT and MC peptides contributed to the separation between the M1–M3 (porcine) and the M4–M6 (bovine) trypsin digests. FT peptides were more abundant in porcine trypsin digests (positive PC2 values), whereas MC peptides were more prevalent in bovine trypsin digests (negative PC2 values). ST peptides contributed much less toward the total intensity despite their high count and were not informative for distinguishing the bovine from the porcine trypsin digests. The peptides that contributed the most toward the separation between bovine and porcine trypsins are labeled in Figure 4A. The most striking example is the MC peptide K.RMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK.C (three missed cleavages) and its shorter MC sibling K.RMPCAEDYLSVVLNQLCVLHEKTPVSDR.V (two missed cleavages), which were favored by the bovine trypsins, whereas the FT subsequence of this peptide, R.MPCAEDYLSVVLNQLCVLHEK.T, was favored in the porcine digests.

The third principal component (PC3; which explained ~2% of the remaining variance) accounted for the differences between the 2 and 18 h digests (Figure 4B). One interesting example is the MC peptide K.RMPCAEDYLSVVLNQLCVLHEK.T,

which was highly abundant in the 2 h porcine digests. This peptide became digested into the FT peptide R.MPCAEDYLSVVLNQLCVLHEK.T at 18 h. Overall, the digestion time effect was notable for the porcine trypsins but largely negligible for bovine trypsins.

To further analyze the peptides, we calculated fold changes ($FC_{\text{bov/por}}$) to identify which peptides were most different in abundance between the bovine and porcine trypsin digests. For each time point (2 and 18 h digestion time), Table 3 lists the top 30 peptides: the 15 peptides with the highest FC values (more abundant in the bovine digests) and the 15 peptides with the lowest FC values (more abundant in the porcine digests). Included in this list are the peptide K.RMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK.C and its subsequences already noted using the PCA analysis described above. At 2 h, Table 3 lists 12 FT and 18 MC peptides, of which all FT peptides were more abundant in the porcine digests and 15 out of 18 MC peptides were more abundant in the bovine digests. A similar trend was observed at 18 h (10 FT peptides, all more abundant in the porcine digests, and 17 MC peptides, two of which were more abundant in the porcine digests), except the list also

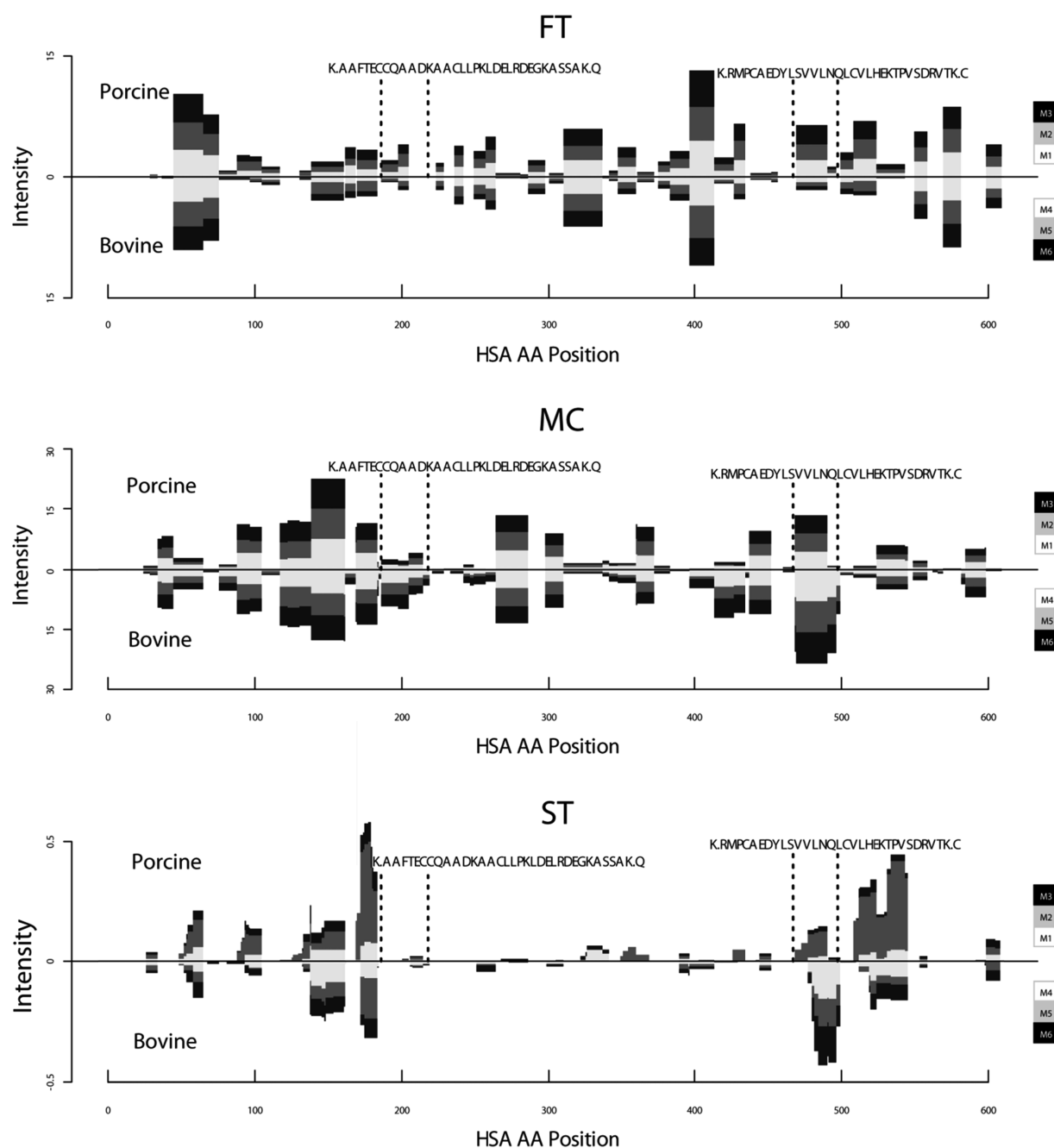


Figure 6. Percent of total intensity contributed by different categories of peptides for each amino acid position. Data shown are for the 18 h time point and are an extension of the analysis in Figure 5A. The shaded colors in each graph represent each manufacturer listed from the lowest number outward: M1–M3 and M4–M6. Porcine peptides are plotted in the positive *y* axis and bovine in the negative *y* axis. The *x* axis represents the amino acid position of HSA, and the *y* axis is the relative % intensity of the total intensity detected at that amino acid position. Examples of sequence regions showing significant differences between porcine and bovine trypsins are indicated. (See the text.) FT: fully tryptic, MC: missed cleavage, ST: semitryptic.

included three ST peptides that were more abundant in the porcine digests. The full list of identified peptides and their fold changes is shown in the Supplementary Tables 3 (2 h digests) and 4 (18 h digests) in the Supporting Information.

Visualization of Trypsin Digestion Using Quantitative “Sequence Maps”

The significant differences between the activities of porcine and bovine trypsins, despite their high sequence identity (82.5%), might be due to their slight structural differences (see structures of porcine (PDB: 2A31) and bovine (PDB: 3MI4) online at

www.rcbs.org). In addition, the denaturing conditions for the digests may produce different changes in the activity or selectivity of the two trypsins for HSA cleavage sites. Here selectivity of trypsin is defined as efficiency of digestion that is dependent on factors such as the relative position of amino acid residues to the K/R cleavage site and the secondary structure of the substrate. This secondary structure could also include structures reproducibly formed in urea. Therefore, we sought to develop a strategy that would allow us to visualize how well each enzyme digested HSA across the HSA sequence. We also sought to determine whether these trends changed over time, dependent on the

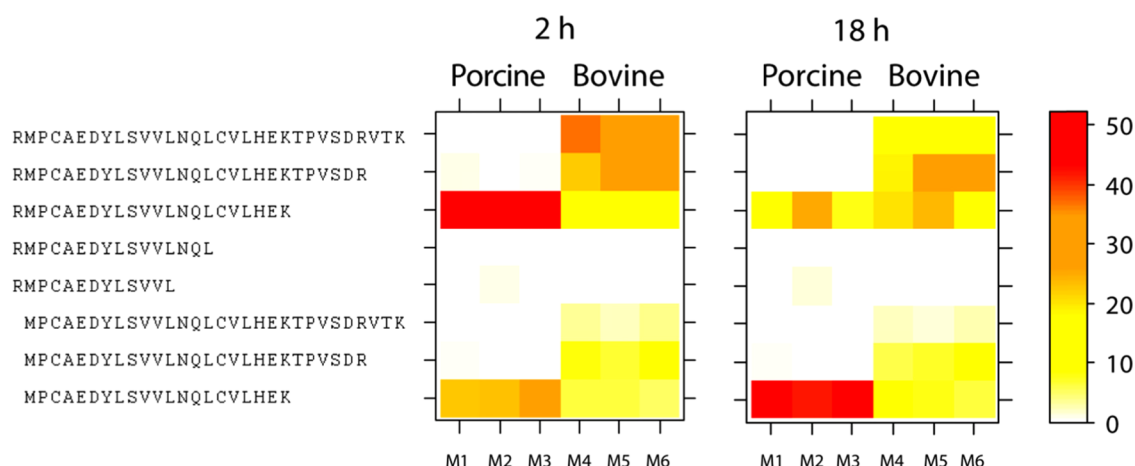


Figure 7. Heat map of peptide abundances. The longest sequence and its subsequences identified in different digests are shown for one exemplary case (see the text for detail). The heatmaps were plotted using the sum of the intensities for each peptide ion that was detected for that sequence. The mean intensity for the three LC–MS/MS replicates were then calculated for each trypsin.

position of the cleavage site in the HSA sequence, and whether these trends were different for the two trypsins.

The intensities of the FT, MC, and ST peptides dependent on the HSA sequence position were plotted in Figure 5. The intensities for each amino acid (AA) position along the HSA sequence were computed as the sum of the intensity of all peptides from a particular category (FT, MC, or ST) at that amino acid position. These intensities were plotted for the 2 (Figure 5A) and 18 h (Figure 5B) digests (for clarity, only the M3 digests are shown; the others responded similarly). At 2 h, the most abundant FT peptide (Figure 5A; AA positions 397–413, peptide K.VFDEFKPLVEEPQNLIK.Q) had an intensity value of 0.0028 (i.e., it contributed 0.28% to the total HSA abundance). The maximum MC peptide intensity, the most abundant peptide overall (position 139–161, missed cleavage site at AA position 160, peptide R.LVRPEVDVMCTAFHD-NEETFLKK.Y), was 0.0046 (see also Figure 4A). The maximum ST peptide intensity was 0.0001 (sequence positions 173–183, peptide Y.FYAPELLFFAK.R). For the 18 h digests, the previously referenced FT peptide increased in intensity to 0.0030, whereas the intensity of the referenced MC peptide remained unchanged. A different peptide, R.MPCAEDYLSVVLNQLCVLHEK.T (AA positions 470–490), was then the most abundant FT peptide (0.0035, an increase from 0.0015 at 2 h). The longer, overlapping peptide, K.RMPCAEDYLSVVLNQLCVLHEK.T (missed cleavage at AA position 469), proportionally decreased from 0.0029 to 0.0005, similar to the results that were observed in the PCA analysis (Figure 4B).

To further compare the differences between the two sequences of trypsin, we plotted the sequence maps of 18 h digests for all of the bovine and porcine trypsins (Figure 6). At 18 h, the total sequence coverage was 99.7 and 99.2% (598aa and 592aa) for the porcine and bovine trypsins. Only two positions in the HSA sequence were not detected in any digest (Q220, R221), whereas three additional HSA amino acids were not detected in any bovine digest (S46Q, S47I, and S48K). This visual representation of the data provides a complementary way to identify sequence regions showing the most significant differences, such as for the KAAFTECCQAADKAACLLPKL-DELRDEGKASSAKQ and KRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK sequences.

The sequence KRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTK and its subsequences are a representative example of

peptides containing missed cleavages (Figure 7) that produced complex digestion dynamics dependent on the length of the digest and the origin of trypsin. For example, in the 2 h digests, both bovine and porcine trypsins were less likely to produce the fully tryptic peptide R.MPCAEDYLSVVLNQLCVLHEK.T, thus favoring the K.R cleavage site over the R.M site (as would be expected according to previous studies⁴⁴). With a longer digestion time, the porcine trypsins become more effective at producing this specific R.M cleavage, whereas the bovine trypsins' ability to cleave at that site did not significantly change. Regardless of the content and relative location of amino acid residues near the cleavage site, digests from bovine trypsins produced nested sequence sets containing more abundant MC peptides.

The differences in the specificities of several commercially available trypsins were also recently observed in the analysis of more complex samples.^{38,39} Because of significantly increased sample complexity, the number and relative intensity of the ST and MC peptides identified in those studies was lower (e.g., ST peptides contributed ~5% of total protein abundance in the analysis of eight-protein mixtures³⁹ vs ~20% in our study). The experimental design used in our work (including lot-to-lot analysis and two digestion time points for each trypsin) and the computational pipeline and visualization tools assembled for the analysis of the data were developed to best ascertain the differences in the efficiency of digestion between different trypsins. Importantly, by using a single protein as the substrate and performing peptide identification using a comprehensive HSA spectral library we were able to identify more MC and ST peptides from the target protein. Additionally, the sample preparation steps in our study did not include fractionation of the digested samples, thus reducing the sources of measurement variability. As a result, our study allowed a more consistent and robust detection of the differences among the ST, MC, and FT peptides when considering the origin of the trypsin. Because we used a single protein as the substrate, we did not attempt to ascertain how digestion may be influenced by the presence of other proteins, as is the case for complex protein samples.³³

CONCLUSIONS

We characterized the performance of six commercially available trypsins, which produced distinct differences depending on the origin (porcine vs bovine). By limiting the substrate to a single

protein (HSA) and digesting it for 2 and 18 h, we were able to gain a more complete picture of the complex digestion process. An initial experiment suggested high reproducibility among digest replicates, whereas the main analysis revealed that porcine and bovine trypsin reproducibly produced a different complement of peptides. Of these, the fully tryptic peptides were observed with the lowest variance of peptide counts and abundance. There were significant differences for how the bovine and porcine trypsin digests produced different sequences and subsequences arising from missed cleavages. The bovine trypsin digests produced a higher number and intensity of MC peptides, whereas porcine trypsin produced more ST peptides. Overall, our analysis suggested that peptide release during the protein digestion depends on multiple factors, including the digestion conditions, the sequence properties of the substrate, and the activity of the protease. Further work is required using well-defined protein substrates such as HSA but also using complex protein samples to obtain a more complete characterization of proteolytic digestion in a typical proteomic experiment. Further studies of the structural relationship between enzyme and substrate together with other causative effects (e.g., conditions for solubility, denatured structure of HSA) to which these differences arise are also needed. The strategies and tools for visualization and data analysis presented in this work, including quantitative sequence maps and PCA, should be useful in these efforts. Further characterization of the complex patterns of trypsin activity and selectivity, together with the improved computational methods for detection of "proteotypic" peptides based on revised trypsin digestion rules, should improve the reliability of peptide detection and quantification. This, in turn, is expected to lead to improved statistical outcomes for proteomic studies involving targeted (e.g., SRM-based) and untargeted experimental workflows.

■ ASSOCIATED CONTENT

■ Supporting Information

Human serum albumin library statistics. Principle components analysis. Identified peptides and normalized intensities from the 2 and 18 h digests. Full list of detected peptide ions for the 2 and 18 h digests. This material is available free of charge via the internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: steve.stein@nist.gov. Tel: +1 301 975 2505.

*E-mail: nesvi@umich.edu. Tel: +1 734 764 3516. Fax: +1 734 936 7361.

Notes

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has been partially funded with NIH grant R01-GM-094231, the National Institute of Standards and Technology

under project "Understanding measuring variability in mass spectrometry based proteomics", and by an Interagency Agreement ACO13004 between NIST and NIH/NCI, and the Clinical Proteomic Tumor Analysis Consortium.

■ REFERENCES

- (1) Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R. Generating and navigating proteome maps using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **2010**, *11* (11), 789–801.
- (2) Cox, J.; Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. In *Annual Review of Biochemistry*; Kornberg, R. D., Raetz, C. R. H., Rothman, J. E., Thorner, J. W., Eds.; Annual Review: Palo Alto, CA, 2011; Vol. 80, pp 273–299.
- (3) Ewles, M.; Goodwin, L. Bioanalytical approaches to analyzing peptides and proteins by LC–MS/MS. *Bioanalysis* **2011**, *3* (12), 1379–1397.
- (4) Manuilov, A. V.; Radziejewski, C. H.; Lee, D. H. Comparability analysis of protein therapeutics by bottom-up LC–MS with stable isotope-tagged reference standards. *mAbs* **2011**, *3* (4), 387–395.
- (5) Olah, T. V.; Slecza, B. G.; D'Arienzo, C.; Tymiak, A. A. Quantitation of therapeutic proteins following direct trypsin digestion of dried blood spot samples and detection by LC–MS-based bioanalytical methods in drug discovery. *Bioanalysis* **2012**, *4* (1), 29–40.
- (6) Bantscheff, M.; Kuster, B. Quantitative mass spectrometry in proteomics. *Anal. Bioanal. Chem.* **2012**, *404* (4), 937–8.
- (7) Addona, T. A.; Abbatiello, S. E.; Schilling, B.; Skates, S. J.; Mani, D. R.; Bunk, D. M.; Spiegelman, C. H.; Zimmerman, L. J.; Ham, A.-J. L.; Keshishian, H.; Hall, S. C.; Allen, S.; Blackman, R. K.; Borchers, C. H.; Buck, C.; Cardasis, H. L.; Cusack, M. P.; Dodder, N. G.; Gibson, B. W.; Held, J. M.; Hiltke, T.; Jackson, A.; Johansen, E. B.; Kinsinger, C. R.; Li, J.; Mesri, M.; Neubert, T. A.; Niles, R. K.; Pulsipher, T. C.; Ransohoff, D.; Rodriguez, H.; Rudnick, P. A.; Smith, D.; Tabb, D. L.; Tegeler, T. J.; Variyath, A. M.; Vega-Montoto, L. J.; Wahlander, A.; Waldemarson, S.; Wang, M.; Whiteaker, J. R.; Zhao, L.; Anderson, N. L.; Fisher, S. J.; Liebler, D. C.; Paulovich, A. G.; Regnier, F. E.; Tempst, P.; Carr, S. A. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **2009**, *27* (7), 633–641.
- (8) Kinsinger, C. R.; Apffel, J.; Baker, M.; Bian, X. P.; Borchers, C. H.; Bradshaw, R.; Brusniak, M. Y.; Chan, D. W.; Deutsch, E. W.; Domon, B.; Gorman, J.; Grimm, R.; Hancock, W.; Hermjakob, H.; Horn, D.; Hunter, C.; Kolar, P.; Kraus, H. J.; Langen, H.; Lindner, R.; Moritz, R. L.; Omenn, G. S.; Orlando, R.; Pandey, A.; Ping, P. P.; Rahbar, A.; Rivers, R.; Seymour, S. L.; Simpson, R. J.; Slotta, D.; Smith, R. D.; Stein, S. E.; Tabb, D. L.; Tagle, D.; Yates, J. R.; Rodriguez, H. Recommendations for Mass Spectrometry Data Quality Metrics for Open Access Data (Corollary to the Amsterdam Principles). *J. Proteome Res.* **2012**, *11* (2), 1412–1419.
- (9) Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. Interlaboratory study characterizing a yeast performance standard for benchmarking LC–MS platform performance. *Mol. Cell. Proteomics* **2010**, *9* (2), 242–254.
- (10) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2010**, *9* (2), 761–776.

- (11) Brownridge, P.; Beynon, R. J. The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods* **2011**, *54* (4), 351–360.
- (12) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012**, *404* (4), 939–965.
- (13) Picotti, P.; Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **2012**, *9* (6), 555–566.
- (14) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* **2012**, *11*, (6).
- (15) Russell, M. R.; Lilley, K. S. Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling. *J. Proteomics* **2012**, *77*, 441–54.
- (16) Xia, J. Q.; Sedransk, N.; Feng, X. Variance component analysis of a multi-site study for the reproducibility of multiple reaction monitoring measurements of peptides in human plasma. *PLoS One* **2011**, *6* (1), e14590.
- (17) Kuhn, E.; Whiteaker, J. R.; Mani, D. R.; Jackson, A. M.; Zhao, L.; Pope, M. E.; Smith, D.; Rivera, K. D.; Anderson, N. L.; Skates, S. J.; Pearson, T. W.; Paulovich, A. G.; Carr, S. A. Interlaboratory Evaluation of Automated, Multiplexed Peptide Immunoaffinity Enrichment Coupled to Multiple Reaction Monitoring Mass Spectrometry for Quantifying Proteins in Plasma. *Mol. Cell. Proteomics* **2012**, *11*, (6).
- (18) Burkhardt, J. M.; Schumbrutski, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **2012**, *75* (4), 1454–1462.
- (19) Olsen, J. V.; Ong, S. E.; Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **2004**, *3* (6), 608–14.
- (20) Keil, B. Proteolysis Data Bank: specificity of alpha-chymotrypsin from computation of protein cleavages. *Protein Sequences Data Anal.* **1987**, *1* (1), 13–20.
- (21) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7* (1), 300–5.
- (22) Rypniewski, W. R.; Perrakis, A.; Vorgias, C. E.; Wilson, K. S. Evolutionary divergence and conservation of trypsin. *Protein Eng.* **1994**, *7* (1), 57–64.
- (23) Swaney, D. L.; Wenger, C. D.; Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **2010**, *9* (3), 1323–9.
- (24) Tran, B. Q.; Hernandez, C.; Waridel, P.; Potts, A.; Barblan, J.; Lisacek, F.; Quadroni, M. Addressing trypsin bias in large scale (phospho)proteome analysis by size exclusion chromatography and secondary digestion of large post-trypsin peptides. *J. Proteome Res.* **2011**, *10* (2), 800–11.
- (25) Vale, G.; Santos, H. M.; Carreira, R. J.; Fonseca, L.; Miró, M.; Cerdà, V.; Reboiro-Jato, M.; Capelo, J. L. An assessment of the ultrasonic probe-based enhancement of protein cleavage with immobilized trypsin. *Proteomics* **2011**, *11* (19), 3866–3876.
- (26) Kim, J.; Kim, B. C.; Lopez-Ferrer, D.; Petritis, K.; Smith, R. D. Nanobiocatalysis for protein digestion in proteomic analysis. *Proteomics* **2010**, *10* (4), 687–699.
- (27) Aguilar-Mahecha, A.; Kuzyk, M. A.; Domanski, D.; Borchers, C. H.; Basik, M. The Effect of Pre-Analytical Variability on the Measurement of MRM-MS-Based Mid- to High-Abundance Plasma Protein Biomarkers and a Panel of Cytokines. *PLoS One* **2012**, *7*, (6).
- (28) Proc, J. L.; Kuzyk, M. A.; Hardie, D. B.; Yang, J.; Smith, D. S.; Jackson, A. M.; Parker, C. E.; Borchers, C. H. A Quantitative Study of the Effects of Chaotropic Agents, Surfactants, and Solvents on the Digestion Efficiency of Human Plasma Proteins by Trypsin. *J. Proteome Res.* **2010**, *9* (10), 5422–5437.
- (29) Bronsema, K. J.; Bischoff, R.; van de Merbel, N. C. Internal standards in the quantitative determination of protein biopharmaceut- icals using liquid chromatography coupled to mass spectrometry. *J. Chromatogr., B* **2012**, *893*, 1–14.
- (30) Freeman, E.; Ivanov, A. R. Proteomics under Pressure: Development of Essential Sample Preparation Techniques in Proteomics Using Ultrahigh Hydrostatic Pressure. *J. Proteome Res.* **2011**, *10* (12), 5536–5546.
- (31) Glatter, T.; Ludwig, C.; Ahrne, E.; Aebersold, R.; Heck, A. J. R.; 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), 5145–5156.
- (32) Hervey, W. J.; Strader, M. B.; Hurst, G. B. Comparison of digestion protocols for microgram quantities of enriched protein samples. *J. Proteome Res.* **2007**, *6* (8), 3054–3061.
- (33) Hustoft, H. K.; Reubsaet, L.; Greibrokk, T.; Lundanes, E.; Malerod, H. Critical assessment of accelerating trypsination methods. *J. Pharm. Biomed. Anal.* **2011**, *56* (5), 1069–1078.
- (34) Klammer, A. A.; MacCoss, M. J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **2006**, *5* (3), 695–700.
- (35) Ren, D.; Pipes, G. D.; Liu, D. J.; Shih, L. Y.; Nichols, A. C.; Treuheit, M. J.; Brems, D. N.; Bondarenko, P. V. An improved trypsin digestion method minimizes digestion-induced modifications on proteins. *Anal. Biochem.* **2009**, *392* (1), 12–21.
- (36) Sun, L. L.; Li, Y. H.; Yang, P.; Zhu, G. J.; Dovichi, N. J. High efficiency and quantitatively reproducible protein digestion by trypsin-immobilized magnetic microspheres. *J. Chromatogr., A* **2012**, *1220*, 68–74.
- (37) Wisniewski, J. R.; Mann, M. Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis. *Anal. Chem.* **2012**, *84* (6), 2631–2637.
- (38) Burkhardt, J. M.; Schumbrutski, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **2012**, *75* (4), 1454–62.
- (39) Bunkenborg, J.; Espadas, G.; Molina, H. Cutting Edge Proteomics: Benchmarking of Six Commercial Trypsins. *J. Proteome Res.* **2013**, *12*, 3631–41.
- (40) Court, M.; Selevsek, N.; Matondo, M.; Allory, Y.; Garin, J.; Masselon, C. D.; Domon, B. Toward a standardized urine proteome analysis methodology. *Proteomics* **2011**, *11* (6), 1160–71.
- (41) Pedrioli, P. G. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol. Biol.* **2010**, *604*, 213–38.
- (42) Hedstrom, L. Serine protease mechanism and specificity. *Chem. Rev.* **2002**, *102* (12), 4501–24.
- (43) Kim, J. S.; Monroe, M. E.; Camp, D. G., 2nd; Smith, R. D.; Qian, W. J. In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J. Proteome Res.* **2013**, *12* (2), 910–6.
- (44) 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.