SUPPLEMENTARY MATERIAL FOR

Confetti: A Multi-protease Map of the *HeLa* Proteome for Comprehensive Proteomics

Xiaofeng Guo, David C. Trudgian, Andrew Lemoff, Sivaramakrishna Yadavalli & Hamid Mirzaei

CONTENTS

SUPPLEMENTARY METHODS

Cell Culture & Lysate Preparation

Human cervical cancer *HeLa* cells were grown in DMEM supplemented with 10% FBS (Invitrogen) in a humidified incubator at 37°C and 5% CO_2. For the experiments, 1 × 10^6 *HeLa* cells were harvested from a 10 cm dish after reaching 80% confluency. Cells were washed twice with phosphate buffered saline (PBS), and resuspended in modified RIPA lysis buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% NP-40, 0.25% Na deoxycholate, 1 mM EDTA) containing a cOmplete Protease Inhibitor Cocktail (Roche Diagnostics, Laval, QC). Cells collected using a scraper were lysed by incubation at 4°C for 10 min with the lysis buffer followed by centrifugation at 14000 rpm for 30 min to pellet the cell debris. Supernatants were collected and proteins were precipitated using trichloroacetic acid at a concentration of 20% w/v on ice for 30 min. After TCA precipitation and centrifugation protein amount was estimated by weighing the dried pellet. Proteins were reduced by incubating at the presence of 8mM Tris(2-carboxyethyl)phosphine (TCEP) at 37°C for 30 min, followed by alkylation with 12mM iodoacetamide (IAA) at 37°C for 30 min in the dark. A final protein concentration of 2mg/ml was achieved in digestion buffer containing 100 mM Tris pH 8.0, 50% trifluoroethanol, 6M urea, 2M thiourea, and 0.5% SDS.

Digest Procedures

In general digest reactions were started by diluting the reaction solution 10-fold using a dilution buffer containing 100mM Tris pH 8.0 and 10mM $CaCl_2$, except for Asp-N and Glu-C digestion reactions, where the solution was diluted 20-fold using either 10mM Tris pH 8.0 (for Asp-N) or 100mM Tris PH 8.0 (for Glu-C) to reduce the final concentration of SDS to 0.025%. Arg-C digestion solutions were diluted 100x using 50mM Tris (pH7.4), 10mM $CaCl_2$, 5mM DTT, and 0.5mM EDTA. Digest timings were chosen so that samples could be prepared in a single working day. All single digestions proceeded for 8 hours using a Thermomixer (Eppendorf) with temperature set at 37°C and mixing speed set at 1400 rpm, except for Asp-N digestion, which was incubated for 16 hours overnight. All double sequential digestions were carried out for 4 hours for the first digestion, followed by another 4 hours for the second digestion. All triple sequential digestions were carried out for 3 hours for the first digestion, followed by 3 hours for the second digestion, and 3 hours for the third digestion. Trypsin, Glu-C, Lys-C, Asp-N, and Arg-C were added to the protein mixture in a 1:50 ratio, whereas elastase and chymotrypsin were added to the protein mixture in a 1:25 ratio. The reactions were stopped by adding TFA to a final concentration of 0.5%. Desalting and SDS removal were accomplished simultaneously using Oasis MCX µElution plates (Waters, Cat# 186001830BA), and the eluate was subsequently dried using a speed vacuum concentrator. In detail, the MCX cartridges were conditioned by passing 0.2ml of methanol, and followed by passing 0.2ml of water containing 0.1% TFA. After passing the samples through the conditioned cartridges, it was washed sequentially by passing 0.2ml of water containing 0.1% TFA, 0.2ml 80% ACN containing 0.1% TFA, and 0.2ml of water. Finally the peptides were eluted with 0.2ml of 90% ACN 10% NH4OH.

LC-MS/MS Methods

All digests were analyzed using Q Exactive and Orbitrap Elite mass-spectrometers (Thermo Fischer, Bremen), coupled to identical Ultimate 3000 RSLCnano HPLC systems (Dionex, Sunnyvale CA). Peptides were loaded onto either a 75 µm i.d. x 50 cm column packed in-house with a reverse-phase material ReproSil-Pur C18-AQ, 1.9µm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany), or a 75 µm i.d. x 50 cm Thermo Scientific Easy-Spray column packed with 2µm resin. A 160 min linear gradient of 1%-41% acetonitrile in 0.1% formic acid followed by a 10min ramp to 80%ACN, using a flow rate of at 400nl/min, was employed to elute peptides from the column. The self-packed column temperature was maintained at 60°C using a butterfly heater (Phoenix S&T, Inc. Chester, PA), and the Easy-Spray column was heated to 60°C using the integrated heater. CID analyses were performed on the Orbitrap Elite instrument with full-MS scans acquired in the Orbitrap at 240K resolution (at m/z 400) followed by 13 data-dependent MS/MS scans acquired by the linear ion trap in the rapid mode. HCD analyses were performed on the Q Exactive instrument using a data-dependent top 20 method, with the full-MS scans acquired at 70K resolution (at m/z 200) and MS/MS scans acquired at 17.5K resolution (at m/z 200). The under-fill ratio was set at 0.3%, with a 3 m/z isolation window and fixed first mass of 100 m/z for the MS/MS acquisitions. For both instruments, charge exclusion was applied to exclude unassigned and charge 1 species, and dynamic exclusion was used with duration of 15 seconds.

Peptide Identification

MGF format peak-lists were generated from RAW data files using ProteoWizard msconvert (version 3.0.3535) (1) . Vendor centroid peak-picking was enabled and MS/MS spectra were filtered to a maximum of 300 most-intense peaks. For each combination of digest, instrument, and fragmentation method, all replicates were analyzed as a single submission using CPFP version 2.0.3 (2, 3). Within CPFP, searches were performed using OMSSA 2.1.8 (4) and X!Tandem 2008.12.01.1 (5) using the k-score plugin (6) and the results were combined using PeptideProphet and iProphet version 4.5.1 (7, 8).

Proteome Amino Acid Coverage (PAAC) Computation & Forward Selection

All peptide spectrum matches at a 1% FDR were imported from CPFP into the Confetti MySQL database. The master peptide to protein mapping and PAAC calculations are implemented in Perl. PAAC as defined above does have the disadvantage that it overestimates true coverage of the proteome, since a peptide present in multiple protein sequences is counted in all of these sequences, even if only one protein was truly present in the sample. An alternative approach counting only non-redundant amino-acid coverage of the proteome would under-estimate coverage, where peptides truly present in the sample from multiple proteins or isoforms are only counted once. We believe the chosen metric provides a good balance between accuracy and speed for the purposes of this study.

The greedy forward-selection algorithm (9) begins by selecting a digest *d1*, which has the highest individual PAAC, as a starting point. The PAAC of *d1* combined with each of the 47 remaining digests is then computed, and the best of the additions selected as *d2* so that *d1* and *d2* are the most complementary pair of digests. The process repeats, adding the most complementary digest remaining at each iteration. In this manner the most complimentary set of 5 digests can be chosen with only 230 PAAC computations. However, the solution is approximate. The algorithm may not select the global optimum combination that can be guaranteed with an exhaustive search.

Protein Inference – Minimal Set Cover

Peptide-to-protein mappings were represented in a graph structure. A protein node is linked to a peptide node by an edge if the protein contains that peptide. The complete graph can be separated into a series of smaller disconnected sub-graphs, since shared peptides tend to cluster – a set of peptides is usually shared amongst a small set of related proteins. Each disconnected sub-graph contains an ambiguous set of protein IDs, with shared or unique peptide evidence. The greedy minimal set-cover algorithm reduces the list of proteins in each sub-graph to a minimal parsimonious set of protein groups. In an iterative procedure the protein ID that explains the greatest number of remaining peptides is selected, until every peptide in the group is covered by selected protein IDs. Each protein ID selected is supported by at least 1 unique peptide ID. The remaining unselected proteins in the graph share all or some of the peptides belonging to selected protein IDs, and are merged into a protein group as sameset or subset entries respectively.

The MSC algorithm was implemented in Perl and performs inference on multiple sub-graphs in parallel using multiple CPU cores, speeding up inference vs single-threaded tools. Further implementation details are provided in supplementary methods. The final algorithm required approximately one hour to perform protein inference on our complete peptide set using a recent 16-core server. Although our data is smaller in number of PSMs than some other large studies, the variety of digests used led to significantly more unique peptide sequences than a typical tryptic or limited-enzyme approach. This severely increases the complexity of protein inference, by introducing more peptide nodes and peptide to protein edges into the graph. The size of the sub-graphs to be solved by the MSC algorithm is greatly increased.

SRM Peptides

Synthetic heavy-isotope containing peptides were obtained from Pierce Biotechnology,(Rockford, IL). These peptides were PEPotec Grade 3, which have typical sequence purities of 50-60% and isotopic purities >99%. All tryptic peptides contained either C-terminal Lys[13C(6)15N(2)] or Arg[13C(6)15N(4) with the exception of NLHYFNSDSFASHPNYPYSDEY which contains a Phe[13C(9)15N(1). AspN peptides are labeled with one of Lys[13C(6)15N(2)], Phe[13C(9)15N(1), Arg[13C(6)15N(4) , Pro[13C(5)15N(1)], or Leu[13C(6)15N(1)]. Heavy amino acids for AspN peptides were selected based on availability, cost, mass shift and proximity to the C-terminal. We selected an amino acid closest to the C-terminal that had a large enough mass shift to prevent cross-talk between heavy and light.

SRM Assay Optimization & Analysis

The top seven transitions for each peptide were determined by monitoring all singly and doubly charged y and b ions below m/z = 1250 for all doubly and triply charged peptide ions below m/z = 1000 for the heavy-labeled peptide standards. In cases where a doubly charged peptide ion had m/z > 1000, the 4+ ion was monitored instead of 2+. These data were analyzed in Skyline v1.4 (http://skyline.maccosslab.org) (10), and collision energies were generated by the software without additional user optimization. Lists of the transitions, retention times, declustering potentials, and collision energies used for tryptic and Asp-N peptides are given in Supplementary Tables S2 and S3.

Spiked cell lysate samples were separated on a Dionex Acclaim PepMap100 reverse-phase C18 column (75 um x 15 cm) using an Ultimate 3000 Nano LC System (Dionex). The HPLC was controlled using Chromeleon Xpress (version 6.8 SR10) and Dionex Chromatography MS Link v. 2.12. Separation of peptides was carried out at 200 nl/min using a gradient from 0-25% B in 15 minutes, 25-35% B in 5 minutes, and 35-80% B in 5 minutes, where mobile phase A was 0.1% formic acid in water and mobile phase B was 0.08% formic acid in 10% water, 80% acetonitrile, and 10% trifluoroethanol. Mass spectrometric analysis was performed on an AB SCIEX (Foster City, CA, USA) 6500 QTRAP mass spectrometer in positive-ion mode running Analyst v.1.6. Scheduled SRM was performed in low-mass mode (m/z = 5-1250) with a target scan time of 2 s and a retention time window of 3 minutes around each peak. SRM data was analyzed using Skyline v1.4.

Criteria for transition selection: Transition chromatograms were manually inspected to ensure presence of sufficient confirming ions, absence of interference, and correct peak selection with reference to signals from the heavy peptide standards. To ensure the presence of sufficient confirming ions and to confirm that the correct peak was selected, seven transitions were monitored for each peptide, with three selected for quantitation. The presence of all seven light transitions at the same retention time as the heavy labeled peptide transitions confirmed that the peak selection was correct. In cases where the light peptide signal was very weak, the closest light peak within 0.2 minutes of the heavy peptide peak was selected. To determine whether interference was present in the transition peaks, the ratio of the peak intensities for heavy relative to light were monitored as was the rank of the peak intensity for light and heavy. Any inconsistent heavy/light ratio or significant difference in peak intensity rank eliminated that transition from consideration for selection for quantitation.

KEY TO SUPPLEMENTARY TABLES

Table S1 – Identification statistics for individual digests and MS analyses. Total and per-replicate PSM and unique peptide counts. Proteome Amino Acid Coverage (PAAC).

Table S2 – Forward selection tree for HCD runs only. Peptide identifications from 48 single/double/triple enzyme digests were iteratively combined to maximize total Proteome Amino Acid Coverage (PAAC) with each addition.

Table S3 – Forward selection tree for CID runs only. Peptide identifications from 48 single/double/triple enzyme digests were iteratively combined to maximize total Proteome Amino Acid Coverage (PAAC) with each addition.

Table S4 – Confetti build statistics for additional subsets of digests and MS runs. PSM, peptide, protein and coverage statistics for 30 subsets of the data acquired in this study. Additional results including preliminary ETD / CID-ETD analysis that was not pursued.

Table S5 – Digest preference for SRM candidate peptides. For each protein identified in the SAX HCD data, the preferred digest generating 3 candidate SRM peptides with highest spectral counts is listed. Proteins selected for the pilot SRM experiment are highlighted.

Table S6 – Gene Ontology terms enriched among proteins that have preferred SRM candidates from each of Trypsin, LysC, AspN, GluC, ArgC-Trypsin digests. Gorilla analysis, filtered to FDR q-value ≤0.05.

Table S7 – Transition list for AspN SRM assays.

Table S8 – Transition list for tryptic SRM assays.

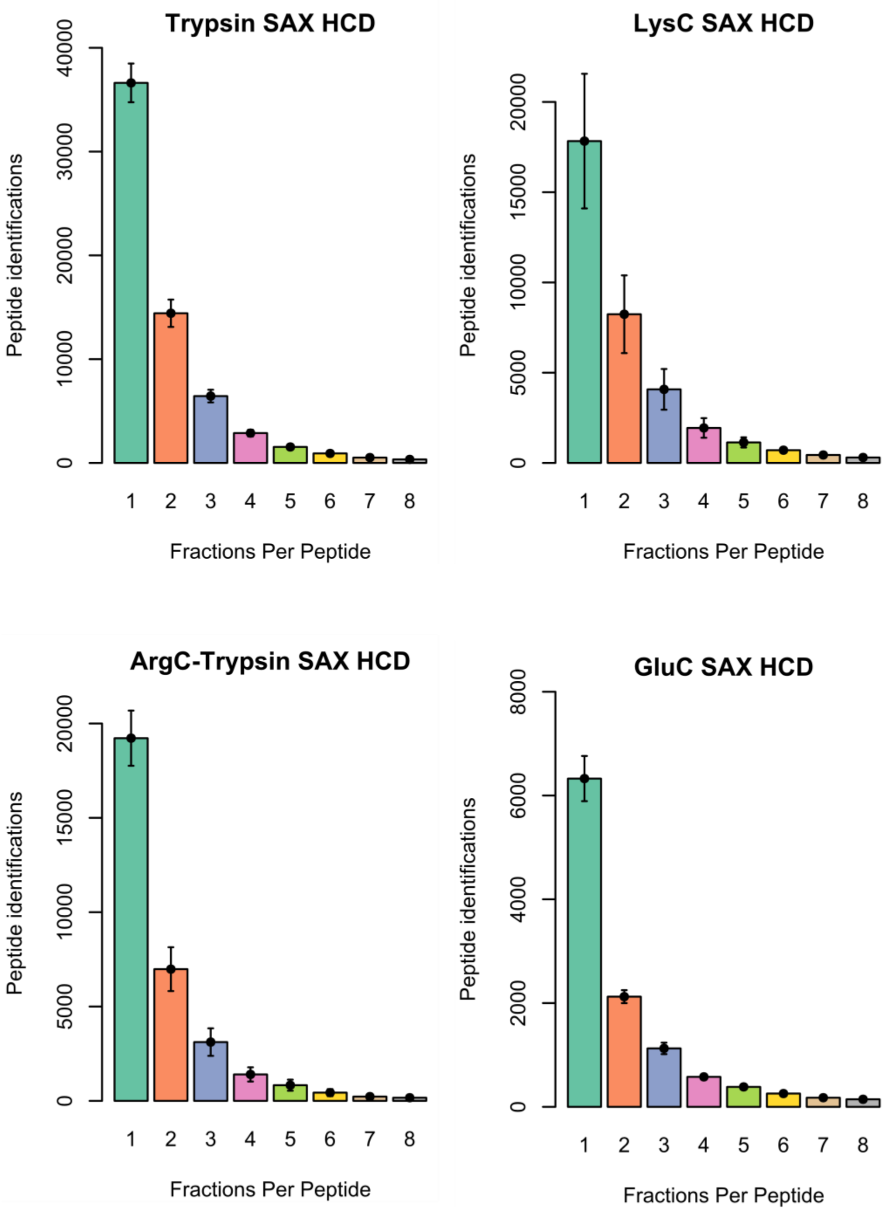
Table S9 – Peak areas for SRM experiment, exported after extraction in Skyline.

Table S10 – List of all protein groups identified in the confetti build.

Table S11 – List of single-peptide protein identifications in the Confetti build, with associated annotated MS/MS spectra.

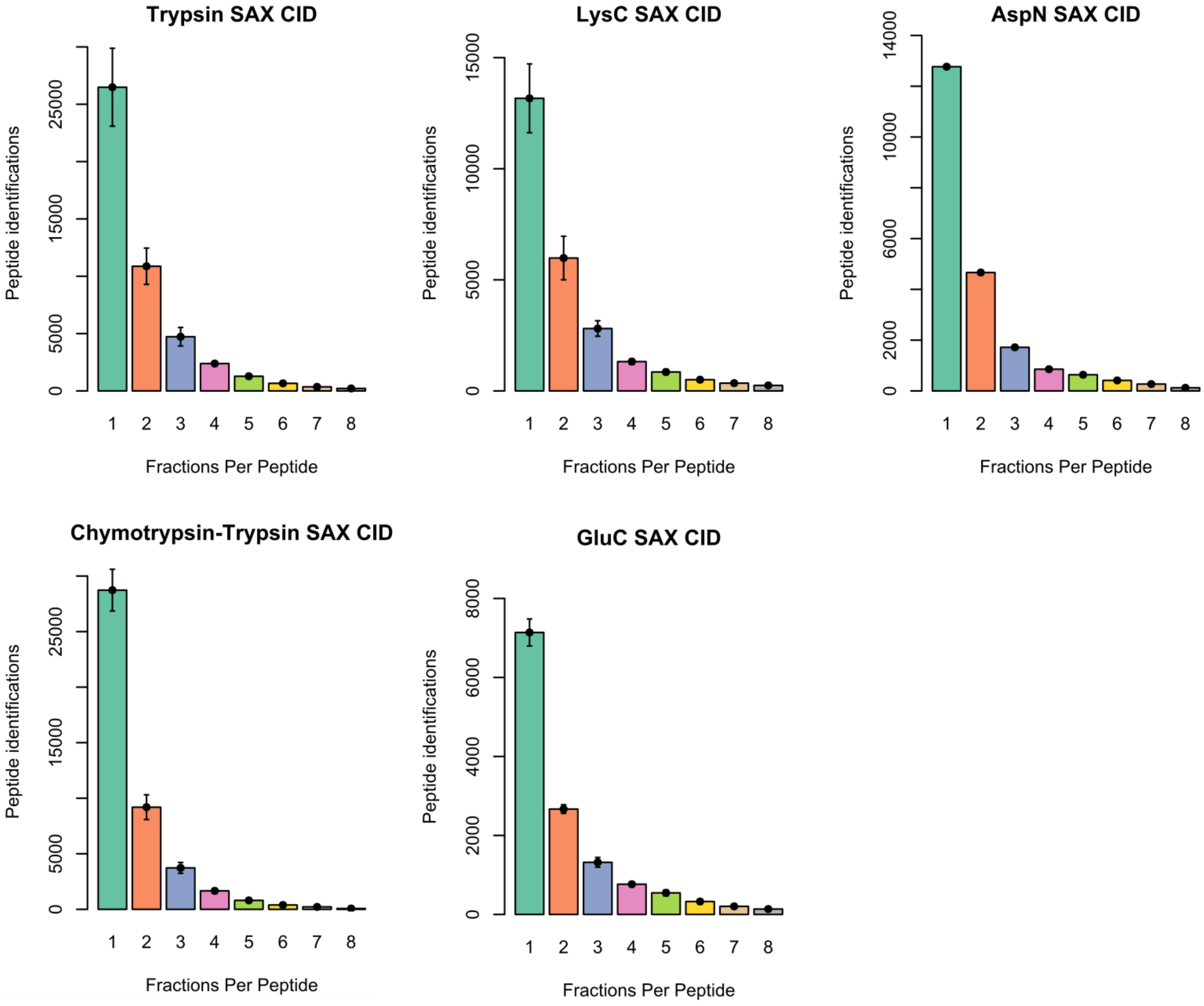
Table S12 – List of peptides / peptide spectrum matches across all datasets.

SUPPLEMENTARY FIGURES

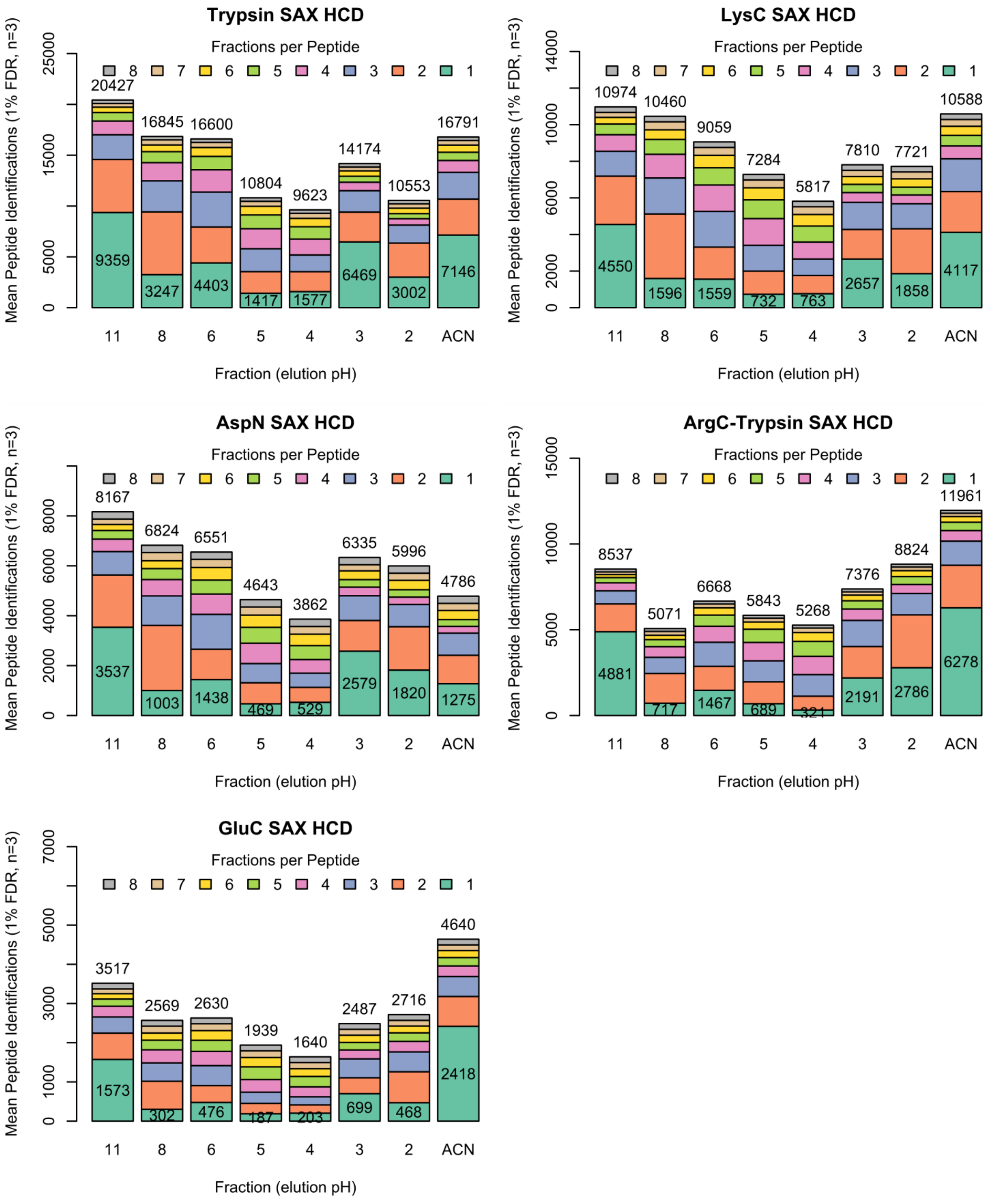


Supplementary Figure S1. Resolution of SAX fractionation for 5 digests analyzed with Q Exactive HCD LC-MS/MS.

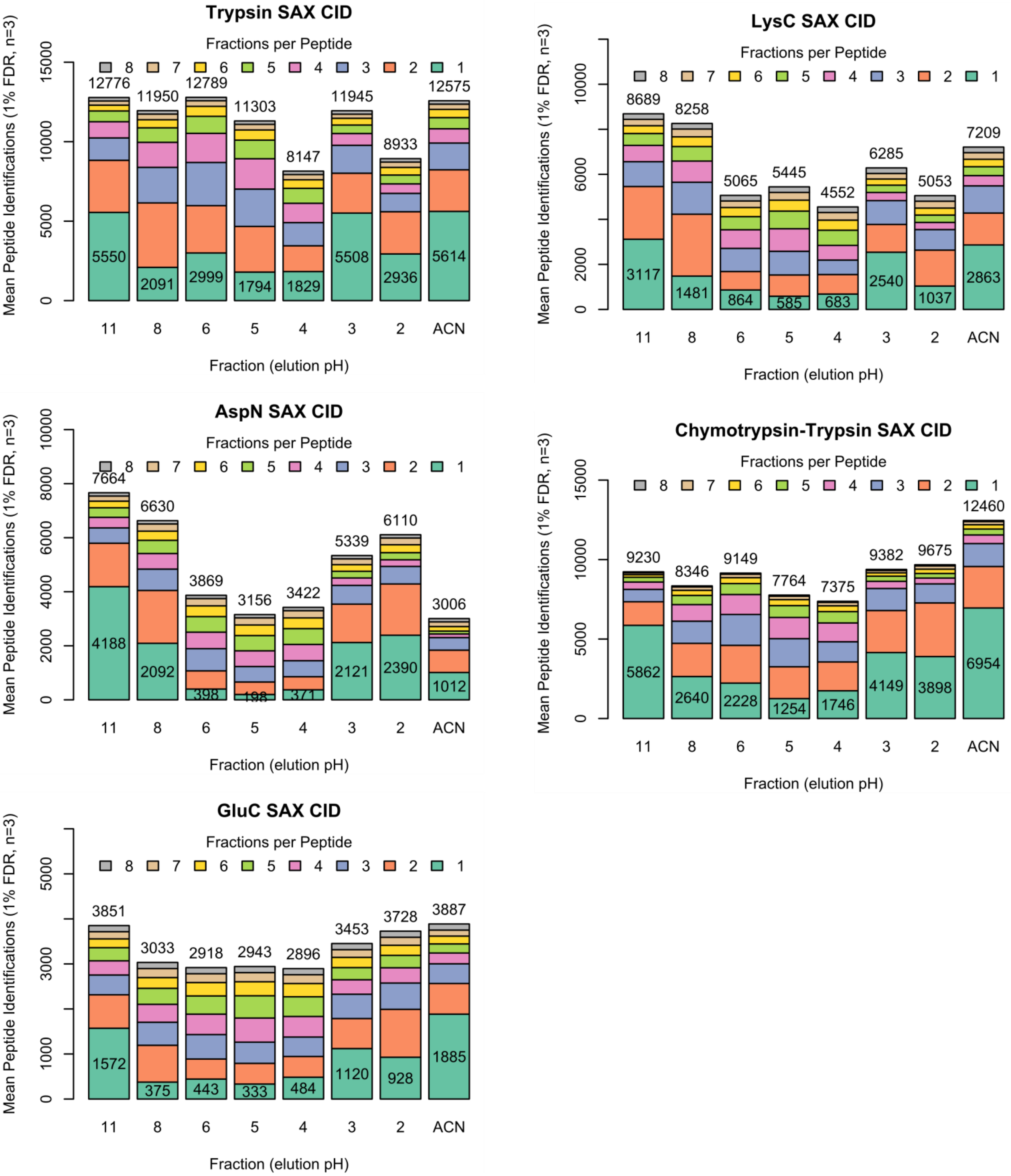
Proportions of fraction-unique and duplicated peptide identifications across the trypsin SAX dataset.



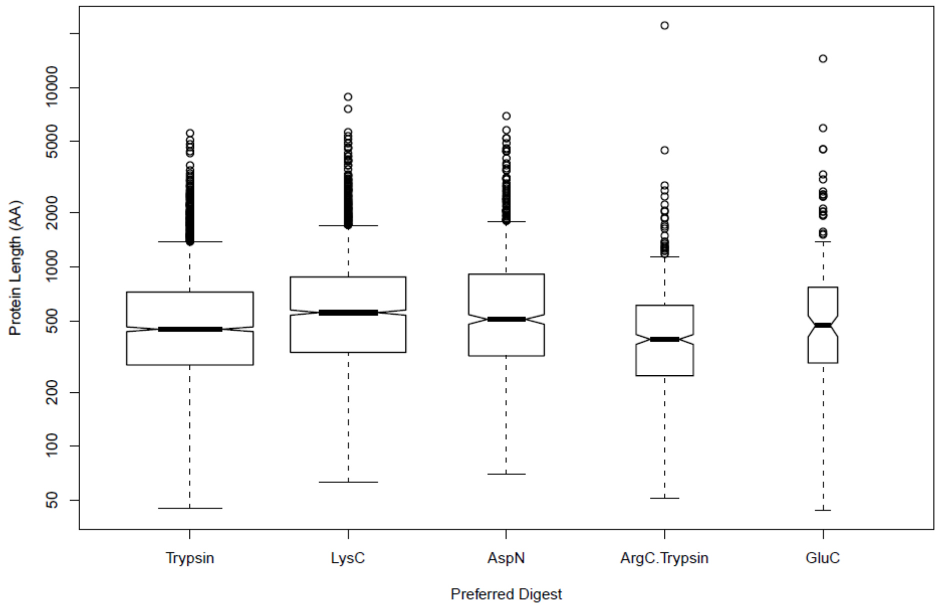
Supplementary Figure S2 - Resolution of SAX fractionation for 5 digests analyzed with Orbitrap Elite CID LC-MS/MS. Proportions of fraction-unique and duplicated peptide identifications across the trypsin SAX dataset.



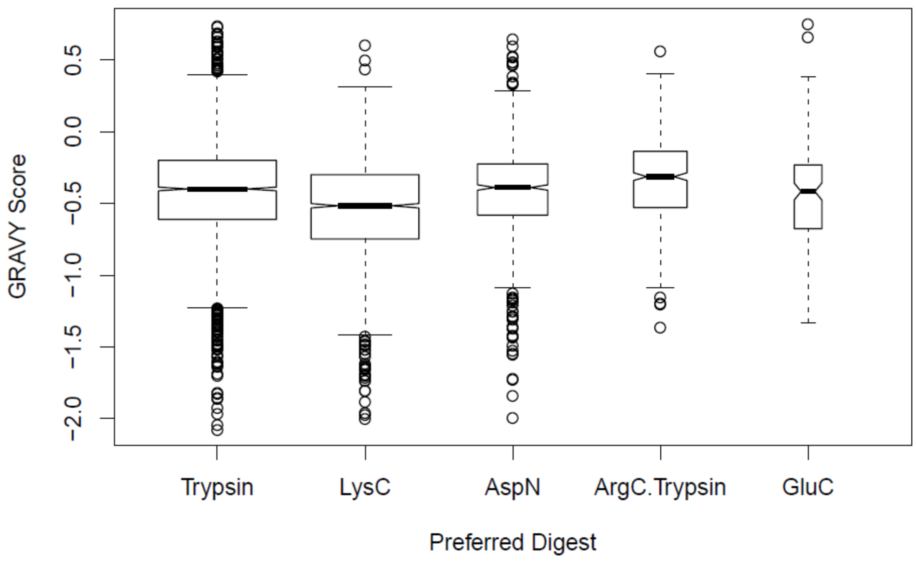
Supplementary Figure S3 - Resolution of SAX fractionation for 5 digests analyzed with Q Exactive HCD LC-MS/MS. Graphs show the mean number of peptide identifications for each fraction, across 3 replicate injections. Colored stacked bars indicate the number of peptides in each fraction that are unique to that fraction (green, labelled) or present in 2-8 fractions. Total peptide IDs per fraction are given above bars.



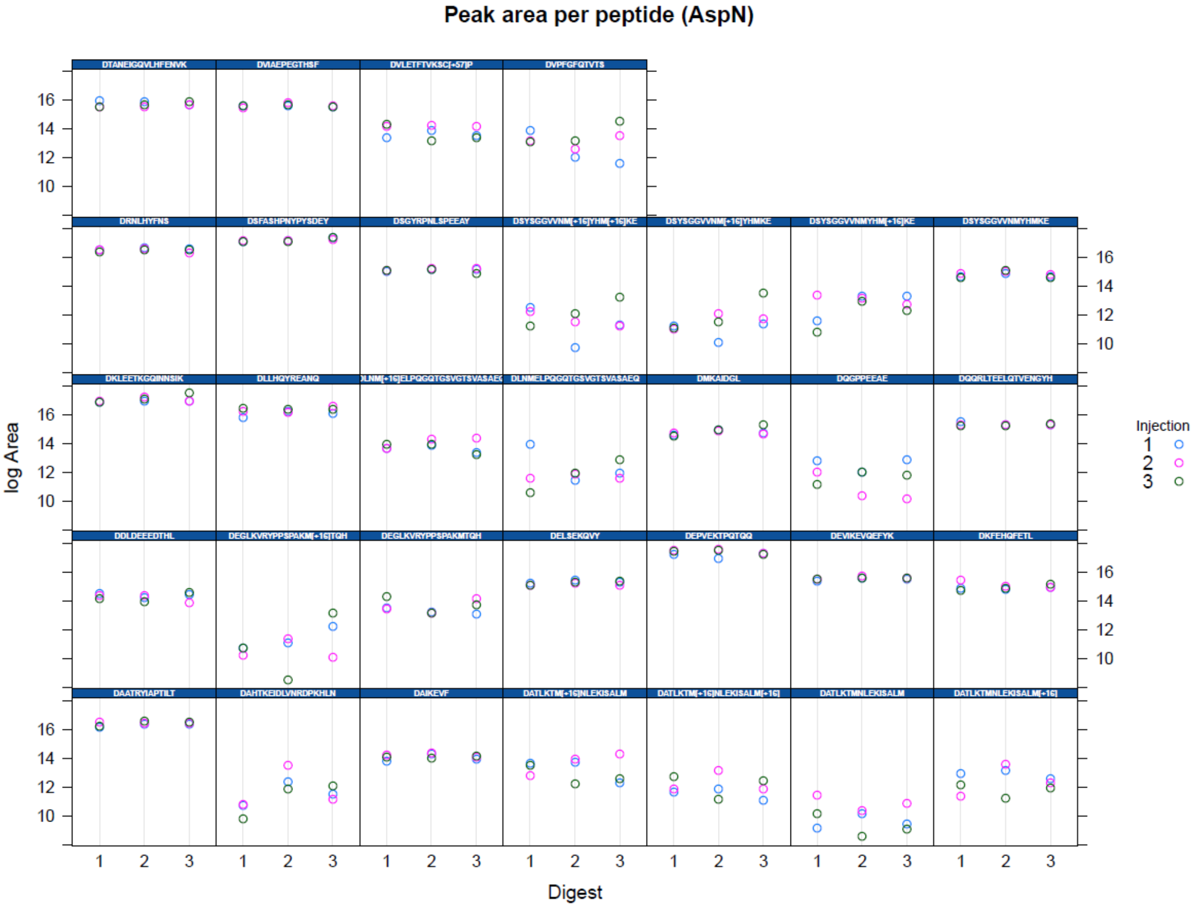
Supplementary Figure S4 - Resolution of SAX fractionation for 5 digests analyzed with Orbitrap Elite CID LC-MS/MS. Graphs show the mean number of peptide identifications for each fraction, across 3 replicate injections. Colored stacked bars indicate the number of peptides in each fraction that are unique to that fraction (green, labelled) or present in 2-8 fractions. Total peptide IDs per fraction are given above bars.



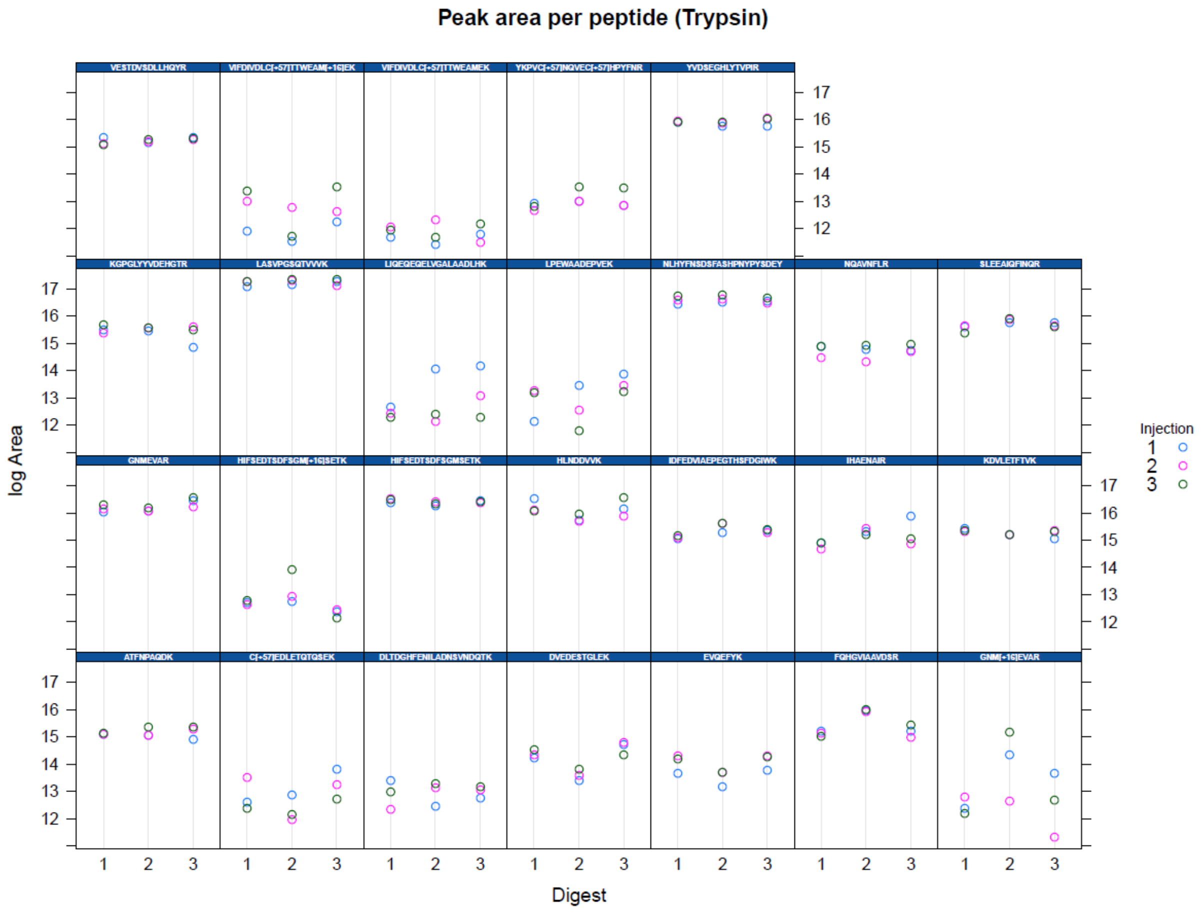
Supplementary Figure S5 – The digest resulting in the set of 3 candidate SRM peptides with highest minimum spectral counts was identified for proteins in the HCD SAX dataset. For the set of proteins assigned to each digest, the distribution of protein lengths was plotted. Boxes show inter-quartile range, with median highlighted. Width of box is proportional to square-root of sample size. Whiskers show +/- 1.5 × interquartile range. Y-axis is log-scale.



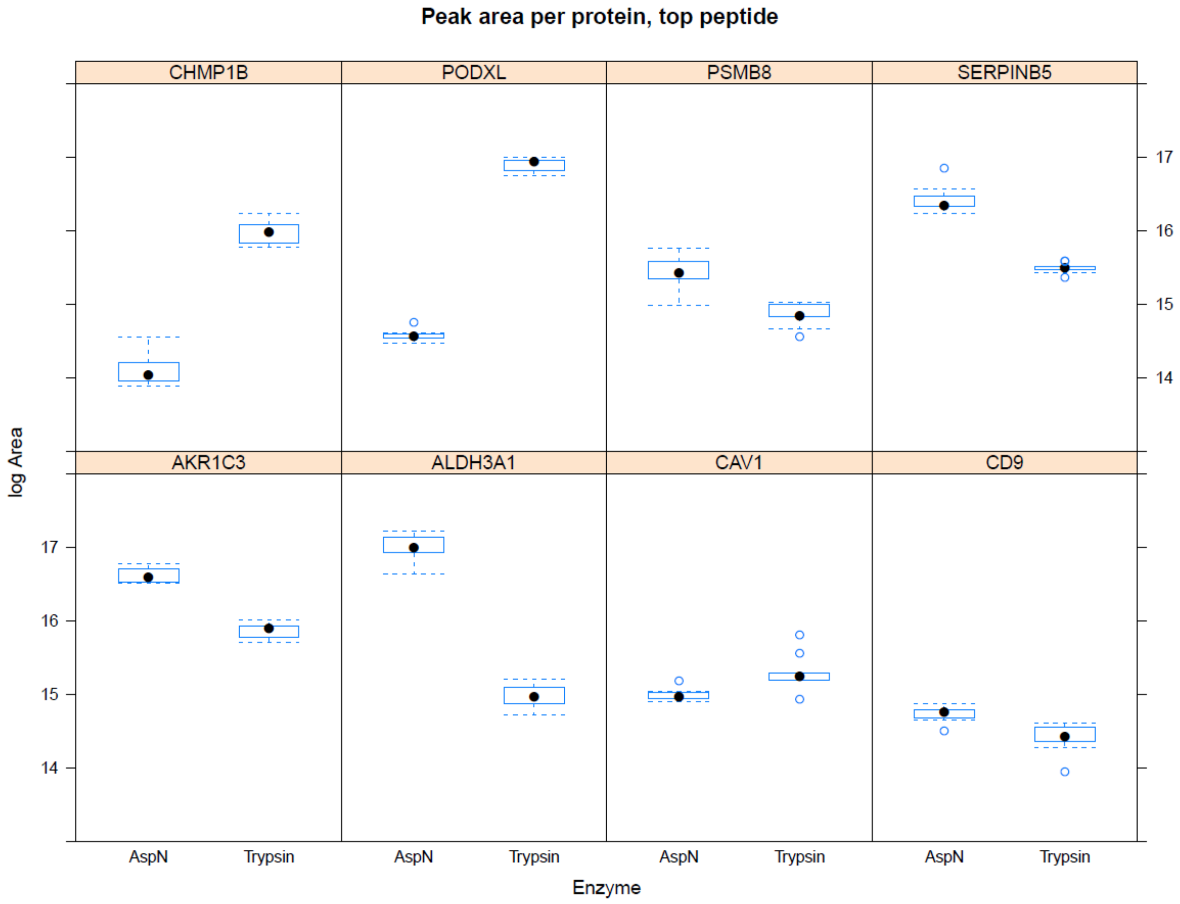
Supplementary Figure S6 – The digest resulting in the set of 3 candidate SRM peptides with highest minimum spectral counts was identified for proteins in the HCD SAX dataset. For the set of proteins assigned to each digest, the distribution of GRAVY score (grand average hydropathy) was plotted. Boxes show inter-quartile range, with median highlighted. Width of box is proportional to square-root of sample size. Whiskers show +/- 1.5 × interquartile range.



Supplementary Figure S7 – Peak areas of SRM transitions for eight proteins in *HeLa* whole-cell lysate, Asp-N digest. Each panel displays replicate injection and digest measurements for a single peptide. The peak areas of the three most intense transitions per peptide were summed. Digest replicates are shown across the panel. Repeat injections for each digest are plotted as colored points.



Supplementary Figure S8 – Peak areas of SRM transitions for eight proteins in *HeLa* whole-cell lysate, tryptic digest. Each panel displays replicate injection and digest measurements for a single peptide. The peak areas of the three most intense transitions per peptide were summed. Digest replicates are shown across the panel. Repeat injections for each digest are plotted as colored points.



Supplementary Figure S9 – Peak areas of SRM transitions for eight proteins in *HeLa* whole-cell lysate, trypsin digest. Each panel displays peak areas per protein, using most intense peptide only.

1. Kessner, D., Chambers, M., Burke, R., Agusand, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536

2. Trudgian, D. C., and Mirzaei, H. (2012) Cloud CPFP: A Shotgun Proteomics Data Analysis Pipeline Using Cloud and High Performance Computing. *Journal of proteome research*

3. Trudgian, D. C., Thomas, B., McGowan, S. J., Kessler, B. M., Salek, M., and Acuto, O. (2010) CPFP: a central proteomics facilities pipeline. *Bioinformatics* 26, 1131-1132

4. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *Journal of proteome research* 3, 958-964

5. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467

6. MacLean, B., Eng, J. K., Beavis, R. C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* 22, 2830-2832

7. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* 74, 5383-5392

8. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* 75, 4646-4658

9. Guyon, I., and Elisseeff, A. (2003) An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3, 1157-1182

10. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966-968