UNIVERSITY OF CALIFORNIA, SAN DIEGO

Novel proteomics methods for increased sensitivity, greater proteome coverage, and global profiling of endogenous SUMO modification sites

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Jesse Gerard Meyer

Committee in charge:

Professor Elizabeth A. Komives, Chair
Professor Nuno Bandeira, Co-Chair
Professor Jack Dixon
Professor Randy Hampton
Professor Judy Kim
Professor Wei Wang

2015

The dissertation of Jesse Gerard Meyer is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

Co-Chair

____

Chair

University of California, San Diego

2015

DEDICATION

I dedicate this work to my family and friends.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACN	Acetonitrile
AGC	Automatic Gain Control
BCA	Bicinchoninic Acid Assay
BEH	Bridged-Ethylene Hybrid
BSA	Bovine Serum Albumin
CID	Collision-induced Dissociation
ddDT	Data-dependent Decision Tree
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
EDTA	Ethylene-diamine Tertiary Acetic Acid
EIC	Extracted Ion Chromatogram
ESI	Electrospray Ionization
ETD	Electron-transfer Dissociation
FA	Formic Acid
FDR	False-discovery Rate
FPLC	Fast Protein Liquid Chromatography
FT	Fourier Transform
FWHM	Full Width at Half Maximum
HCD	Higher-energy Collisional Dissociation
HPRP	High-pH Reversed Phase
IAA	Iodoacetamide

LC	Liquid Chromatography
LC/MS	Liquid Chromatography-Mass Spectrometry
LTQ	Linear Trap Quadrupole
m-NBA	meta-Nitrobenzyl Alcohol
MaLP	M190A Alpha-lytic Protease
mRNA	Messenger Ribonucleic Acid
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MudPIT	Multi-dimensional Protein Identification Technology
NEM	N-ethyl Maleimide
NRAAS	Non-redundant Amino Acid Sequences
PAGE	Poly-acrylimide Gel Electrophoresis
PTM	Post-translational Modification
RNA	Ribonucleic Acid
SDC	Sodium Deoxycholate
SDS	Sodium Dodecyl Sulfate
SL	Sodium Laurate
SUMO	Small Ubiquitin-like Modifier protein
TFA	Trifluoroacetic Acid
TCEP	Tris(2-carboxyethyl)phosphine
TIC	Total-ion Chromatogram
TOF	Time of Flight

TPP	Trans-Proteomic Pipeline
WaLP	Wild-type Alpha-lytic Protease

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many people have helped me accomplish this work. First, I want to thank my family for all their support. My parents have always expected the best from me. I am also thankful to my Aunt Marie helping me transition to life in San Diego.

I could not have completed this work without the help of my advisor, Betsy Komives, and my co-advisor, Nuno Bandeira.

Finally, I want to thank my all my past and present colleagues. I am thankful for help from the following people outside the Komives lab: Sangtae Kim, Xinning Jiang, Judith Coppinger, Eric Bennett, and Lisa Rising. I am thankful for help from the following people from within the Komives lab: Mela Mulvihill, Brian Fugelstad, Nick Treuheit, Morten Beck Trelle, Lindsey Handley, Jimmy Marion, Jonathon Parnell, and Min Wu.

Chapter II, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in *Journal of the American Society for Mass Spectrometry*. (**Meyer, J.G.** and Komives, E. A. (2012). Charge State Coalescence During Electrospray Ionization Improves Peptide Identification by Tandem Mass Spectrometry. Journal of the American Society for Mass Spectrometry, **23,** 1390-1399.)

Chapter III, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in *Molecular and Cellular Proteomics*. (**Meyer, J.G.**, Kim, S., Maltby, D., Ghassemian, M., Bandeira, N., Komives E.A.

(2014). Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. Molecular & Cellular Proteomics, **13**, 823-835.)

Chapter IV, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in *ISRN Computational Biology*. (**Meyer, J.G.** (2014). In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. ISRN Computational Biology, **2014,** 1-7.)

VITA

2009 Bachelor of Science, Biochemistry, University of Minnesota – Twin Cities, Minneapolis, MN

2012 Master of Science, Chemistry, University of California, San Diego

2015 Doctor of Philosophy, Chemistry, University of California, San Diego

PUBLICATIONS

Meyer, J.G. and Komives, E. A. (2012). Charge State Coalescence During Electrospray Ionization Improves Peptide Identification by Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry,* **23**, 1390-1399.

Meyer, J.G. (2014). In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. *ISRN Computational Biology,* **2014**, 1-7.

Meyer, J.G., Kim, S., Maltby, D., Ghassemian, M., Bandeira, N., Komives E.A. (2014). Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. *Molecular & Cellular Proteomics,* **13**, 823-835.

FIELDS OF STUDY

Major Field: Biochemistry
Studies in Biochemistry and Biophysics
Professor Elizabeth A. Komives

HONORS AND AWARDS

2011-2013 Interfaces Graduate Training Program, trainee

2009 Undergraduate Student Poster Award, 23rd Annual Protein Society Symposium

2008 Undergraduate Research Opportunities Program Grant, University of Minnesota

2003 Eagle Scout, Boyscouts of America, Troop 263, Lakeville, MN

ABSTRACT OF THE DISSERTATION

Novel proteomics methods for increased sensitivity, greater proteome coverage, and global profiling of endogenous SUMO modification sites

by

Jesse Gerard Meyer

Doctor of Philosophy in Chemistry

University of California, San Diego, 2015

Professor Elizabeth A. Komives, Chair

Professor Nuno Bandeira, Co-Chair

Proteomics, or the measurement of all proteins present in a biological system under defined conditions, is a relatively young field that is rapidly developing. Currently the best method to achieve high proteome coverage is with bottom-up proteomics, in which the proteome is digested into peptides that are identified followed by inference of their protein origin. Several steps in the bottom-up proteomics workflow leave room for improvement, especially proteome digestion. This work investigates novel bottom-up proteomics methods for improved sensitivity, proteome coverage, and ultimately PTM detection.

Chapter II investigates the effect of two chemicals on peptide electrospray sensitivity. We postulated that peptide supercharging combined with ETD would improve the identification efficiency of peptides, especially nontryptic peptides. We measured the charge state distributions, the total signal, and the number of identified peptides for peptides produced from trypsin, elastase, or pepsin digestion. Unexpectedly, the results show that addition of 5% DMSO to mobile phases used for peptide separation with online ESI resulted in charge state coalescence of peptide signal towards a single charge state, therefore improving signal to noise.

In Chapter III the novel application of two proteases, WaLP and MaLP, for proteome digestion are explored. The results show that the combination of data from separate proteome digestion with trypsin, LysC, WaLP, and MaLP double the observed proteome sequence coverage. The increased coverage was most beneficial for coverage for protein sequences containing too many or too few tryptic cleavage sites. The increased coverage was also beneficial for coverage of proteins with many transmembrane helices.

Chapter IV presents a computational study that attempts to optimize proteome digestion using various real and theoretical cleavage agents. Individual digestions and iterative digestion strategies were simulated. One conclusion of this work is that the

greatest proteome coverage can be obtained using iterative digestion with cleavage starting at the rarest residues first.

Chapter V demonstrates a novel method for untargeted, site-level identification of endogenous SUMO attachment sites in the human proteome. When proteins modified by SUMO are digested with WaLP, a SUMO-remnant diglycyl-lysine modification is left at the site of SUMOylation, which is then detected by tandem mass spectrometry. The results demonstrate identification of 707 unique SUMO modification sites in 443 proteins, of which 414 are previously unknown SUMOylation sites.

Chapter I

Introduction

A. Bottom-up proteomics

After the human genome was sequenced, we learned that the DNA variation among individuals is small (Conrad et al. 2011), and therefore our DNA alone cannot explain the phenotypic differences we observe in our population. Since then, biologists have sought system-wide measurement of all RNAs, proteins, and metabolites. So-called “omics” technologies, once matured, promise the ability to accurately model and predict biology, which has far-reaching implications. Microarray technology/nucleic acid sequencing and mass spectrometry-based proteomics have enabled extensive characterization of RNAs (transcriptome) and proteins (proteome), respectively. Interestingly, protein levels do not entirely correlate with RNA levels (Ghaemmaghami et al. 2003). The complement of proteins in a biological system is more descriptive of the biological state than the complement of mRNA or DNA. Thus, the field of proteomics has developed rapidly over the last decade.

Currently the best way to measure a large number of proteins from a biological system is with shotgun, or bottom-up, proteomics (Walther and Mann 2010), which is a multi-step workflow (Figure 1.1). First, proteins are isolated from cells or tissue using denaturants, such as SDS. Then proteins are digested into peptides with a protease, usually trypsin, which cleaves after positively charged residues arginine and lysine. Those peptides are then subject to one or more dimensions of separation with liquid chromatography, the last of which is coupled to ESI for introduction of the

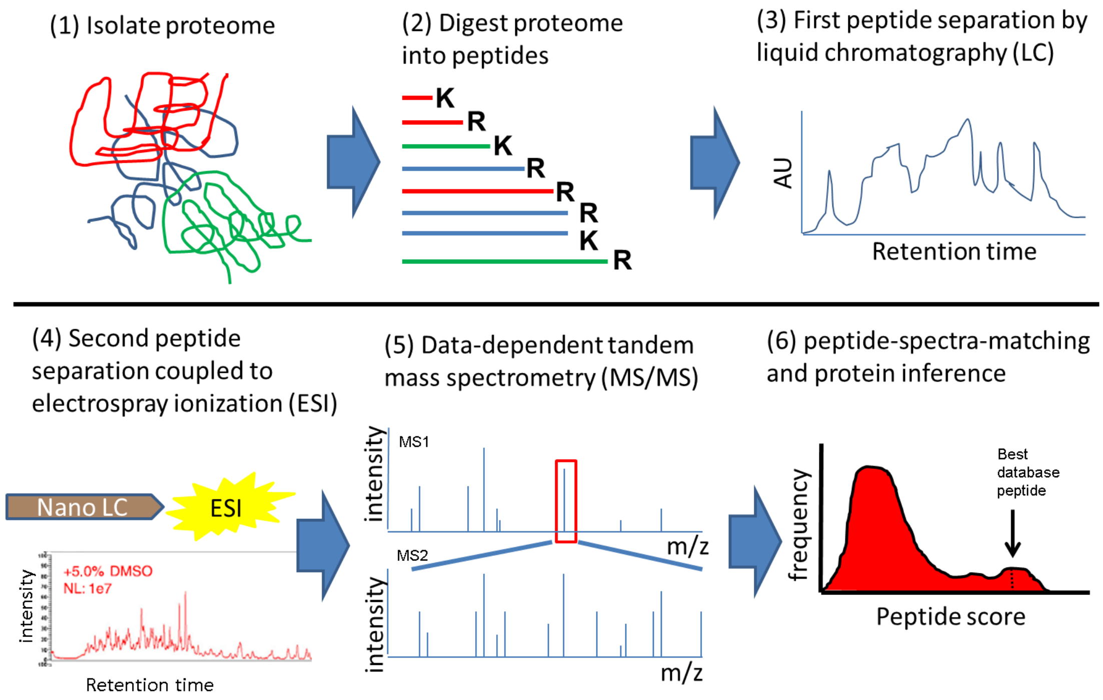

(1) Isolate proteome
(2) Digest proteome into peptides
K
R
K
R
R
R
K
R
(3) First peptide separation by liquid chromatography (LC)
AU
Retention time
(4) Second peptide separation coupled to electrospray ionization (ESI)
Nano LC
ESI
+5.0% DMSO
NL: 1e7
intensity
Retention time
(5) Data-dependent tandem mass spectrometry (MS/MS)
MS1
intensity
m/z
MS2
intensity
m/z
(6) peptide-spectra-matching and protein inference
Best database peptide
frequency
Peptide score

Figure 1.1. The bottom-up proteomics workflow is a multi-step process.

peptides into a MS. The MS then continuously surveys the population of peptide masses that enter the instrument and choses the most abundant signals for fragmentation. The resulting MS/MS spectra reveal specific fragments of the parent peptide that can be matched to peptide sequences. The protein origin is inferred if the spectrum uniquely matches a protein sequence predicted from the genome.

Proteome measurement is a more complex analytical challenge than nucleic acid sequencing for several reasons (Reinders et al. 2004; Zhang et al. 2013). Chemical diversity of proteins, large dynamic range of protein concentrations, and lack of signal amplification all hinder our ability to accurately characterize the entire proteome. Additionally, due to mRNA splice variants and post-translational modifications (PTMs), the number of unique protein sequences that humans express may reach the order of one million.

Every step of the shotgun proteomics workflow has been extensively researched, and great progress has been made. A breakthrough method was introduced in 2001, multidimensional protein identification technology (MudPIT), which enabled identification of over 1,000 proteins from a single sample injection (Washburn et al. 2001). In 2008, Matthias Mann’s group was able to identify and quantify all proteins predicted to be expressed by yeast (de Godoy et al. 2008). These researchers used a large amount of input protein, extensive pre-fractionation, and roughly 40 days of mass spectrometry data acquisition to achieve this. Then in 2010, Joshua Coon’s group explored the use of several proteases for proteome digestion, and found that the combination of data from five separate proteome digestions resulted in over 25% proteome sequence coverage (Swaney et al. 2010). Only two years later, in

2012, Matthais Mann's group was able to identify nearly the entire yeast proteome in a single 4 hour run (Nagaraj et al. 2012). More recently, in 2014, the Coon group was able to further improve their workflow to allow identification of nearly all proteins in yeast in only 70 minutes (Hebert et al. 2014).

Despite the great progress towards complete proteomics, the observed protein sequence coverage is often very low; many proteins are identified by a single peptide sequence. High protein sequence coverage is needed for comprehensive mapping of post-translational modifications (PTMs), and observation of all mRNA splice variants. Therefore, additional research is needed to achieve the lofty goal of complete proteome sequence coverage. Towards the goal of complete proteome sequence coverage, I have developed novel methods focused on novel digestion strategies that increase observed amino acid coverage and reveal previously recondite PTMs (Meyer and Komives 2012; Meyer 2014; Meyer et al. 2014).

B. Peptide Electrospray and Fragmentation

In order to detect a peptide by MS, it must become ionized in the gas phase. This is commonly achieved using ESI. Electrospray of peptides is a complex process where several variables compete to produce a population of charged peptides (Ogorzalek Loo et al. 2014). The charge states of peptides produced during ESI effect the ability to fragment and identify them. Peptides from trypsin digestion are almost exclusively ionized as +2 charge state ions upon ESI, but non-tryptic peptides are likely to have more diverse charge states. Data dependent mass spectrometry methods almost always exclude singly charged peptides from selection for tandem mass spectrometry. Charge states over +5 are also often excluded. Therefore, non-tryptic peptides that lack a basic residue and ionize as +1 ions, or peptides that contain many basic residues and become highly charged upon ESI will not be identified with traditional mass spectrometry methods.

The choice of fragmentation method used should be paired with the expected charge state of peptides. Three ion activation methods are commonly used for peptide fragmentation, CID, ETD, and HCD (Figure 1.2). Each fragmentation method has different effectiveness with different charge densities. CID is more efficient for identification of low charge density precursor ions, but ETD is more efficient for identification of high charge density precursor ions, regardless of the peptide sequence character (see Figure 3.4).

Because non-tryptic are likely to have less defined charge than tryptic peptides upon ESI, any experiment that aims to use non-tryptic peptides should take charge states into consideration. Within the last 15 years, several researchers have found that they can manipulate the charge state of analytes by adding certain chemicals to electrosprayed solutions, e.g. glycerol or DMSO (Iavarone and Williams 2002; Iavarone and Williams 2003). In 2007, one group reported that supercharging from m-NBA combined with ETD resulted in better ETD spectra and improved identifications (Kjeldsen et al. 2007). Based on their findings, and the fact that non-trytic peptides may have low charge states, we postulated that the combination of peptide supercharging and ETD would be beneficial for identification of non-tryptic peptides (Meyer and Komives 2012). In Chapter II, I present an assessment of benefits resulting from supercharging reagents for peptide electrospray. We find that

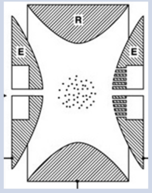
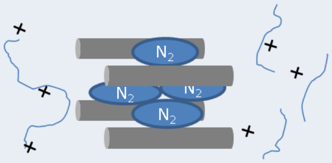
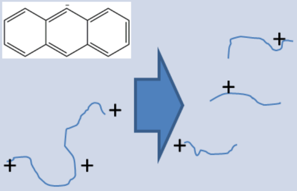
Ion activation	+	-
Collision-induced dissociation (CID) – peptide heated by collisions with inert gas; breaks weakest bonds first , ~30 V input energy  *http://aemc.jpl.nasa.gov/instruments/vcam/*	- Fast - Well characterized fragmentation - Well suited for +2 charge state peptides, e.g. tryptic peptides	- loss of weak covalent modifications (e.g., phosphate/sugar) - Usually low-resolution
Higher-energy collisional dissociation (HCD) – "beam-type" dissociation, kV energy input 	- Some phosphate retention - Produces immonium ions - Usually high resolution (TOF)	- Loss of weak covalent modification, e.g. glycosylation - May produce internal ions
Electron-transfer dissociation (ETD)– gas phase ion-ion reaction between peptide cation and reagent anion 	- Fragment ion series less dependent on peptide sequence composition - Retains labile modifications, e.g. glycosylation	- Slow - Inefficient fragmentation, especially with low precursor charge density - Less accessible, newer method

Figure 1.2. A table describing the ion activation methods CID, HCD, and ETD. The column "+" gives some benefits of each method, and the column "-" gives some drawbacks of each method.

the presence of DMSO during peptide electrospray increases the sensitivity of peptide identification, and we explore several possible reasons for this observation.

C. Proteome digestion

Observable proteome coverage is ultimately limited by the digestion method used to generate peptides. Peptides that are too long cannot be identified for several analytical reasons, and peptides that are too short to uniquely match a protein sequence are also lost. Using *in silico* digestion, we can examine the theoretical length distribution that would be produced from any protease with strict specificity (see Figure 3.1). Most peptides produced from any protease fall below the generally useful lower length limit of 7 amino acids. Using these theoretical length distributions, we can compute theoretical maximum proteome coverage for each digest (Figure 1.3). Even though the combination of 5 proteases is predicted to produce nearly 95% proteome coverage, the observed proteome coverage in a recent study that combined data from proteome digestion with these 5 proteases was only 25% (Swaney et al. 2010). However, the use of separate digestions increases analysis time. Therefore, new proteome digestion strategies are needed that can allow further increases in proteome coverage in less analysis time. Most proteases currently used for proteome digestion cleave after similar charged residues. For this reason, we sought to apply new proteases with substrate specificity that drastically differs from currently employed proteases. We reasoned that cleavage of very different residues could improve proteome coverage more efficiently than the use of several proteases with similar specificity (Meyer et al. 2014). In Chapter III, we demonstrate the novel

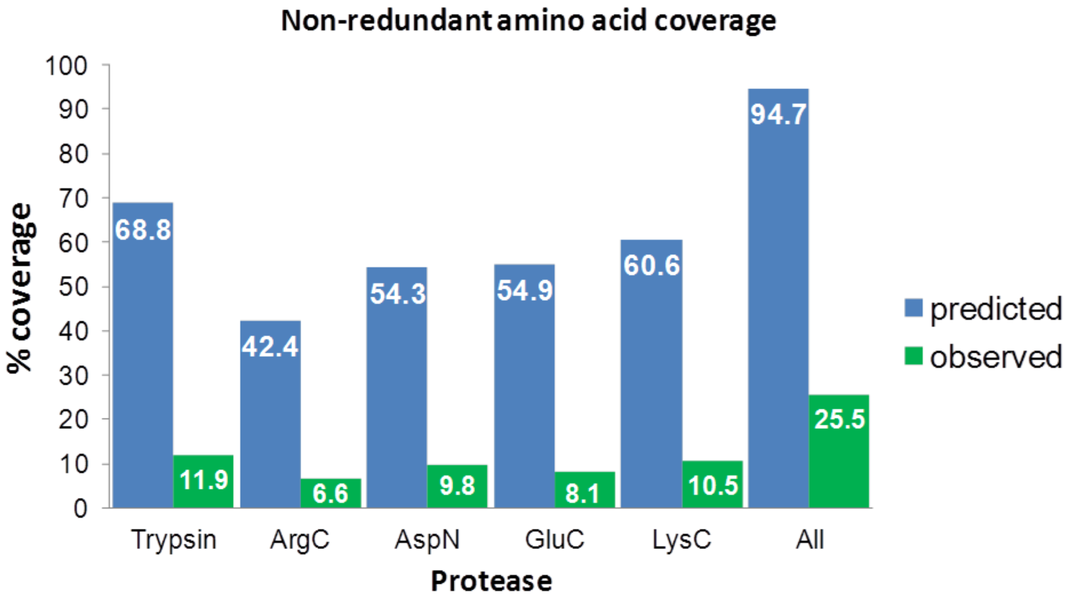


Figure 1.3. A plot comparing the theoretical upper limits of proteome coverage with observed coverage. Observed coverage values were taken from the 2010 paper by Swaney, et al.

application of two proteases to proteome digestion. We explore the benefits of increased proteome coverage resulting from digestion with these new proteases, and we also explore challenges that must be addressed to effectively identify non-tryptic peptides. In Chapter IV, I present an exploration of various digestion agents and the resulting sequence coverage at the whole proteome level and at the residue level. I identify a iterative digestion strategy that theoretically allows very high proteome coverage as compared with any single digestion alone. Finally, in Chapter V, we demonstrate a novel digestion strategy that enables site-level identification of endogenous SUMO modification sites in the human proteome, which has not previously been possible in high throughput.

D. References

Conrad, D. F., J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles and P. Awadalla (2011). "Variation in genome-wide mutation rates within and between human families." Nat. Genet. 43(7): 712-714.

de Godoy, L. M. F., J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther and M. Mann (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature 455(7217): 1251-1254.

Ghaemmaghami, S., W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea and J. S. Weissman (2003). "Global analysis of protein expression in yeast." Nature 425: 737-741.

Hebert, A. S., A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall and J. J. Coon (2014). "The One Hour Yeast Proteome." Molecular & Cellular Proteomics : MCP 13: 339-347.

Iavarone, A. T. and E. R. Williams (2002). "Supercharging in electrospray ionization: effects on signal and charge." Int J Mass Spectrom 219(1): 63-72.

Iavarone, A. T. and E. R. Williams (2003). "Mechanism of Charging and Supercharging Molecules in Electrospray Ionization. ." J Am Chem Soc 125(8): 2319-2327.

Kjeldsen, F., A. M. B. Giessing, C. R. Ingrell and O. N. Jensen (2007). "Peptide Sequencing and Characterization of Post-Translational Modifications by Enhanced Ion-Charging and Liquid Chromatography Electron-Transfer Dissociation Tandem Mass Spectrometry." Anal Chem 79(24): 9243-9252.

Meyer, J. G. (2014). "*In Silico* Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage." ISRN Comp Biol 2014(1): 1-7.

Meyer, J. G., S. Kim, D. A. Maltby, M. Ghassemian, N. Bandeira and E. A. Komives (2014). "Expanding proteome coverage with orthogonal-specificity α-lytic proteases." Mol Cell Proteomics 13(3): 823-835.

Meyer, J. G. and E. A. Komives (2012). "Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry." J Am Soc Mass Spectrom 23(8): 1390-1399.

Nagaraj, N., N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm and M. Mann (2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap." Mol Cell Proteomics 11(3): M111.013722.

Ogorzalek Loo, R., R. Lakshmanan and J. Loo (2014). "What Protein Charging (and Supercharging) Reveal about the Mechanism of Electrospray Ionization." Journal of the American Society for Mass Spectrometry 25(10): 1675-1693.

Reinders, J., U. Lewandrowski, J. Moebius, Y. Wagner and A. Sickmann (2004). "Challenges in mass spectrometry-based proteomics." Proteomics 4(12): 3686-3703.

Swaney, D. L., C. D. Wenger and J. J. Coon (2010). "Value of using multiple proteases for large-scale mass spectrometry-based proteomics." J Proteom Res 9(3): 1323-1329.

Walther, T. C. and M. Mann (2010). "Mass spectrometry-based proteomics in cell biology." J Cell Biol 190(4): 491-500.

Washburn, M. P., D. Wolters and J. R. Yates (2001). "Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology." Nature Biotech 19(3): 242-247.

Zhang, Y., B. R. Fonslow, B. Shan, M.-C. Baek and J. R. Yates (2013). "Protein Analysis by Shotgun/Bottom-up Proteomics." Chemical Reviews 113: 2343-2394.

Chapter II

Charge State Coalescence during Electrospray Ionization Improves Peptide Identification by Tandem Mass Spectrometry

A. Introduction

Tandem mass spectrometry (MS/MS) is a robust, fast and sensitive analytical method that has transformed the field of proteomics (Sleno and Volmer 2004; Ong and Mann 2005). “Shotgun” or “bottom-up” proteomics involves isolation of the entire protein complement of a biological system followed by digestion into smaller fragments with a protease. Peptides are then subject to several dimensions of separation by liquid chromatography (LC) (Motoyama and III 2008), the last of which is coupled directly to electrospray-ionization (ESI) MS/MS (Fenn et al. 1989). Shotgun proteomics has been used to quantify the entire complement of proteins expressed by yeast (de Godoy et al. 2008).

Peptides for proteomic analysis are often generated exclusively by trypsin, because it produces high peptide yield and has high specificity for the positively charged amino acid residues arginine and lysine (Burkhart et al. 2011). Tryptic peptides have desirable length and charge characteristics for identification by collision induced dissociation (CID), which produces *b*- and *y*- ion series (Tabb et al. 2003). Digestion with trypsin alone, however, only covers 11.9% of the non-redundant amino acid sequences (NRAAS) (Swaney et al. 2010). By using a combination of five commercially available proteases, coverage of NRAAS increased to 25.5% (Swaney et al. 2010), which highlights the value of using alternative proteases even though they produce suboptimal peptides for traditional MS/MS identification.

Electron-transfer dissociation (ETD) is a complementary fragmentation method that forms *c*- and *z*- ion fragments in a manner that is less dependent on the peptide sequence (Syka et al. 2004). However, ETD is inefficient with low intensity and/or low charge state precursors. Although the ETD efficiency improves when combined with a resonance excitation after the ion-ion reaction (supplemental activation) (Swaney et al. 2006), the ability to supercharge precursors is expected to dramatically improve ETD efficiencies.

The charge state and intensity of peptide ions is determined at the point of ESI by the competition of several variables including: instrument settings (Loo et al. 1988), chemical properties of the analyte and mobile phase (e.g. peptide pI, mobile phase pH) (Kamel et al. 1999), ion suppression of co-eluting analytes (Mallet et al. 2004), and mobile phase flow rate (Ficarro et al. 2009). Long before ESI was applied to macromolecules, Lord Raleigh predicted that the maximum extent of charging, z_R, possible for a spherical droplet of radius, R, correlates to the surface tension, γ, of the liquid according to the relationship (Rayleigh 1882):

$$\text{Total charge} = Z_R e = 8\pi(\varepsilon_0 \gamma R^3)^{1/2}$$

Where e and ε_0 are constants pertaining to the elementary charge and the permittivity of free space, respectively. Several groups have identified molecules that result in “supercharging” of polymers and intact proteins. Iavarone et al. first demonstrated the charge enhancement of *meta*-nitrobenzyl alcohol (*m*-NBA) for intact proteins (Iavarone and Williams 2002; Iavarone and Williams 2003). The charge enhancement of *m*-NBA and other supercharging reagents during ESI is especially complementary to ETD fragmentation. Addition of 0.1% *m*-NBA was shown to enhance charging and therefore ETD fragmentation of peptides derived from BSA and β-casein (Kjeldsen et al. 2007). Addition of *m*-NBA improved top down H/D exchange with ETD of protein structure yielding 1.3 amino acid resolution in real time (Sterling et al. 2010) and top-down ESI with electron-capture dissociation (ECD) was also improved (Yin and Loo 2011). Additional supercharging reagents have been identified. In 2002, Iavarone et al. reported the use of several small molecules, including dimethylsulfoxide (DMSO) and even glycerol (Iavarone and Williams 2002; Iavarone and Williams 2003). Recently, Valeja et al. reported several small molecule organic reagents that effect ESI charge state of intact protein and chromatographic retention during LC/MS (Valeja et al. 2010). Lomeli et al. also report the extent of charge enhancement resulting from a screen of several aromatic compounds and sulfolane (Lomeli et al. 2010). These researchers reported that sulfolane forms adducts in the supercharging process, and sulfolane adducts were recently investigated in more detail (Douglass and Venter 2012). Despite extensive application of supercharging reagents for ESI of intact proteins, so far only Kjeldsen et al. applied *m*-NBA to enhance the charge state of peptides for ETD fragmentation and identification (Kjeldsen et al. 2007).

Although the charge enhancing phenomena of these reagents correlates well with the Raleigh equation in many measurements, the complex chemical environment at the end of the droplet lifetime produces deviations from theory (Hogan and Biswas 2008; Wilm 2011). Factors affecting the observed analyte charging largely result from the high boiling supercharging reagent that becomes enriched in the late stage of the droplet resulting in: non-spherical droplets (Ahadi and Konermann 2011), protein chemical and/or thermal denaturation (Sterling et al. 2010), and gas-phase proton affinity (Loo and Smith 1995).

Here we present an exploration of charge-enhancing reagents, DMSO and *m*-NBA, for improved peptide identification by LC-MS/MS. DMSO was selected because Valeja et al. recently observed, in addition to supercharging, improved chromatography of intact protein during reversed-phase with a C5 stationary phase (Valeja et al. 2010). *m*-NBA was selected because Kjeldsen et al. previously reported increased ETD quality with this co-solvent (Kjeldsen et al. 2007). Sulfolane was not used since we already had a representative sulfoxide compound (Douglass and Venter 2012). We were particularly interested in the effect of charge enhancement for peptides generated from alternative proteases that do not have an amino acid with a basic side chain at the C-terminus. We hypothesized that high charge states produced from ESI with DMSO and *m*-NBA may improve the sensitivity of non-tryptic peptide identification when used in combination with ETD. To assess the practical application of these reagents for LC-MS/MS, the mobile phases used for reversed-phase nano liquid chromatography were modified with supercharging reagents DMSO and *m*-NBA in a manner similar to Kjeldsen et al (Kjeldsen et al. 2007). Mobile phase properties are known to effect chromatographic resolution and retention (Gustavsson et al. 2001; Coulier et al. 2006), and indeed some differences were observed in the chromatography. Peptides produced by various protease digestions (i.e. trypsin, elastase or pepsin) of a five protein mixture were used to assess effects of the modified mobile-phases on chromatographic separation, precursor ion effects, and ultimately, the number of peptide identifications. The samples were analyzed with a combination of CID and ETD (Swaney et al. 2008; Kim et al. 2011) and the data were searched with MS-GFDB which allows searching of CID/ETD pairs (Kim et al. 2010). The co-solvent, DMSO, markedly improved the number of peptide identifications compared to formic acid alone for all three proteases.

B. Materials and Methods

1. Samples and Solutions

Acetonitrile (ACN) and formic acid (FA) Optima grade were purchased from Thermo Fisher Scientific (Waltham, MA). TCEP (BondBreaker) was from Pierce (Rockford, IL). Trizma, Iodoacetamide (IAA), sodium deoxycholate (SDC) dimethylsulfoxide (DMSO), and meta-nitrobenzyl alcohol (m-NBA), and angiotensin I were purchased from Sigma Aldrich (Saint Louis, MO). All chemicals were of the highest purity possible and were used without further purification. Peptides from a standard mixture of five proteins (Bovine serum albumin (BSA), α-casein (S1 and S2), β-casein, lysozyme C, and hemoglobin (alpha and beta)) were digested separately by pepsin, elastase or trypsin as described previously (Lin et al. 2008), with the exception that pepsin digestion was performed in 0.1% FA without sodium deoxycholate. The digests were stored lyophilized at -80° C. For MS/MS experiments, the amount injected was 16 nanograms. Full scan MS experiments were performed with three quantities of analyte spanning an order of magnitude, 16, 80, and 160 nanograms.

2. Liquid Chromatography

Control mobile phase A consisted of 0.2% FA with 5% ACN in water, and control mobile phase B consisted of 0.2% FA with 95% ACN in water. In addition to the controls, both mobile phases, A and B, were also modified with either 5% DMSO or 0.1% m-NBA. The presence of 5% ACN in mobile phase A assisted with m-NBA dissolution.

3. nLC-MS/MS

Lyophilized peptides were resuspended in 0.2 % FA in water, and were separated over a 75 µm ID X 12 cm capillary column packed in-house with 5 µm Phenomenex Luna C18 particles. An Agilent 1200 series pump was used to generate a flow of 0.11 mL/min, which was split 1:500 to ~250 nL/min. Separation was achieved with a gradient from 100% to 70% A over 60 minutes. Mobile phase B was then increased to 95% over 10 min, and the column was flushed for 10 minutes, followed by re-equilibration with 100% A at 400 nL/min for 10 minutes. MS data were collected during the flush and re-equilibration to ensure consistent column regeneration, resulting in a total of 90 minutes of data collection for each run. The eluent was directly electrosprayed at 2.1 kV into an LTQ XL with ETD (Thermo Scientific, Waltham, MA). To ensure effective desolvation of high-boiling supercharging reagents, the ion transfer tube temperature was set to 275 °C. Full-scan-only experiments were collected from m/z 350-2000 using the enhanced scan rate, which produces resolution sufficient to resolve the isotope distribution of a +3 charge state ion. For data-dependent peptide identification experiments, the top five most abundant precursor ions determined by a precursor ion scan from m/z 300-2000 were selected for “zoom” scans (+/- 5 daltons) at a resolution sufficient to determine precursor charge state. Charge states not equal to 1 with intensity over 1000 counts were fragmented consecutively by both CID and ETD. After collection of two fragment ion spectra, precursor m/z values were excluded for 15 seconds. The exclusion list size was set to 500. Activation settings were optimized by directly infusing angiotensin I at 1 picomole/microliter. Infusion in each of the three different solvent systems did not show any significant differences in the optimal activation settings. ETD activation was performed using fluoranthene to generate anions for an ion/ion reaction time of 120 ms followed by supplemental activation to break up non-covalent gas phase interactions (Swaney et al. 2006). CID activation was performed for 30 ms using “35% normalized collision energy”. The automatic gain control settings were AGC Reagent=3e5, AGC Full MS=3e4, AGC MSN=1e3, AGC zoom=3e3. The instrument was operated using the Xcalibur version 2.0.7 software (ThermoFisher Scientific).

4. Data Analysis

The resulting data was analyzed by MS-GFDB (Kim et al. 2010). Files were converted to mzXML using Trans-Proteomic Pipeline (Keller et al. 2005). The resulting .mzXML files were searched against the Uniprot database of all bovine proteins plus common contaminants and lysozyme C from chicken, as well as the three proteases, porcine trypsin, elastase and pepsin. The false discovery rate (FDR) was estimated using the target-decoy approach by including shuffled sequences in the search database. Database searches were performed with and without the option to merge CID and ETD fragment ion spectra from the same precursor ion (Kim et al. 2010). In addition to the default fixed carbamidomethylation of cysteine, searches allowed variable deamidation of Q or N, and phosphorylation of S or T. Up to two variable modifications were allowed. Data from elastase and pepsin experiments were searched using “no enzyme” specificity. Precursor charge states from two to five were considered for spectra with undetermined charge. The precursor mass tolerance was set to 2.5 daltons. Default parameters used were: instrument= low-resolution LCQ/LTQ; one allowed non-enzymatic terminus, and possible peptide lengths from six to forty amino acids were considered. Only peptide spectrum matches to the standard proteins or the protease with FDR < 0.01 were used for further analysis. Chromatographic retention and resolution were assessed using XCMS (Smith et al. 2006) and in-house scripts written in R (Team 2011). Annotated MS/MS spectra were visualized using Proteowizard (Kessner et al. 2008). Spectral counts, referring to the number of times a peptide sequence was matched to an MS/MS spectrum, were used to provide an estimate of MS/MS efficiency.

C. RESULTS

1. Co-solvent effects on the number of high quality peptide identifications

To explore the improvements in peptide identifications from the use of mobile-phase additives (i.e. 5% DMSO, 0.1% m-NBA), we analyzed digests of a 5 protein mixture containing BSA, hemoglobin, α-casein, β-casein, and lysozyme C. The mixture was digested with trypsin, elastase, or pepsin, and analyzed in triplicate by nLC-MS/MS on an LTQ mass spectrometer using consecutive activation of selected precursors by both CID and ETD fragmentation. The resulting spectra were searched as pairs. At a peptide-level false discovery rate of <0.01, the numbers of unique peptides identified under each solvent condition are given in Table 2.1. For each of the three different proteolytic digests, inclusion of DMSO increased the numbers of peptides identified, and for two of the three protease digestions, the results were highly significant (p-value > 0.05). The observed improvement could arise from a number of variables, such as ESI charge enhancement resulting in more efficient ETD fragmentation, or effects on chromatographic retention and resolution. The contribution of each variable was assessed separately.

2. Effects of co-solvents on chromatography

Most of the data acquisition time during the peptide identification experiments is spent determining charge state and collecting fragment ion spectra (> 5 s between precursor scans). To better assess the contribution of chromatographic quality on the relative MS/MS efficiencies observed, we collected full-scan-only MS spectra. Normalized total ion chromatograms (TICs) are shown in Figure 2.1. Separations containing m-NBA resulted in more noise, as evidenced by the elevated baseline, as well as an apparent increase in peak height and width. Peak heights also apparently increased when the mobile phase contained 5% DMSO. Quantitative assessment of peptide-level chromatographic resolution was achieved by generating extracted-ion chromatograms (EICs), an example of which is shown in Figure 2.2. The full width at half maximum (FWHM) from analyses carried out in FA only and FA plus DMSO runs were not significantly different. However, separations carried out in the presence of m-NBA resulted in wider peaks (Table 2.2). Figure 2.2 also shows how the retention time decreases due to the co-solvents. This observation prompted us to carry out a non-linear retention time alignment as a comprehensive measure of chromatographic retention (Figure 2.3). Separations in the presence of DMSO and m-

Table 2.1. Effect of mobile phase additive on the numbers of unique peptides identified from digests of the 5-protein mixture.

	FA	**+ DMSO**	**+ m-NBA**	**p-value$_1$***	**p-value$_2$****
Trypsin	183 ± 9[1]	231 ± 4	166 ± 6	0.0020	0.0345
Elastase	204 ± 29	242 ± 21	189 ± 43	0.0706	0.3195
Pepsin	372 ± 13	409 ± 7	267 ± 24	0.0108	0.0029

[1] Unique peptide counts are the average of three technical replicates. Error values given are one standard deviation. * P-value from a student's t-test comparing FA alone to FA + DMSO. ** P-value from a student's t-test comparing FA alone to FA + m-NBA. The DMSO modified phase identifies significantly more peptides at p-value <0.05 and <0.01 for peptic and tryptic digests, respectively.

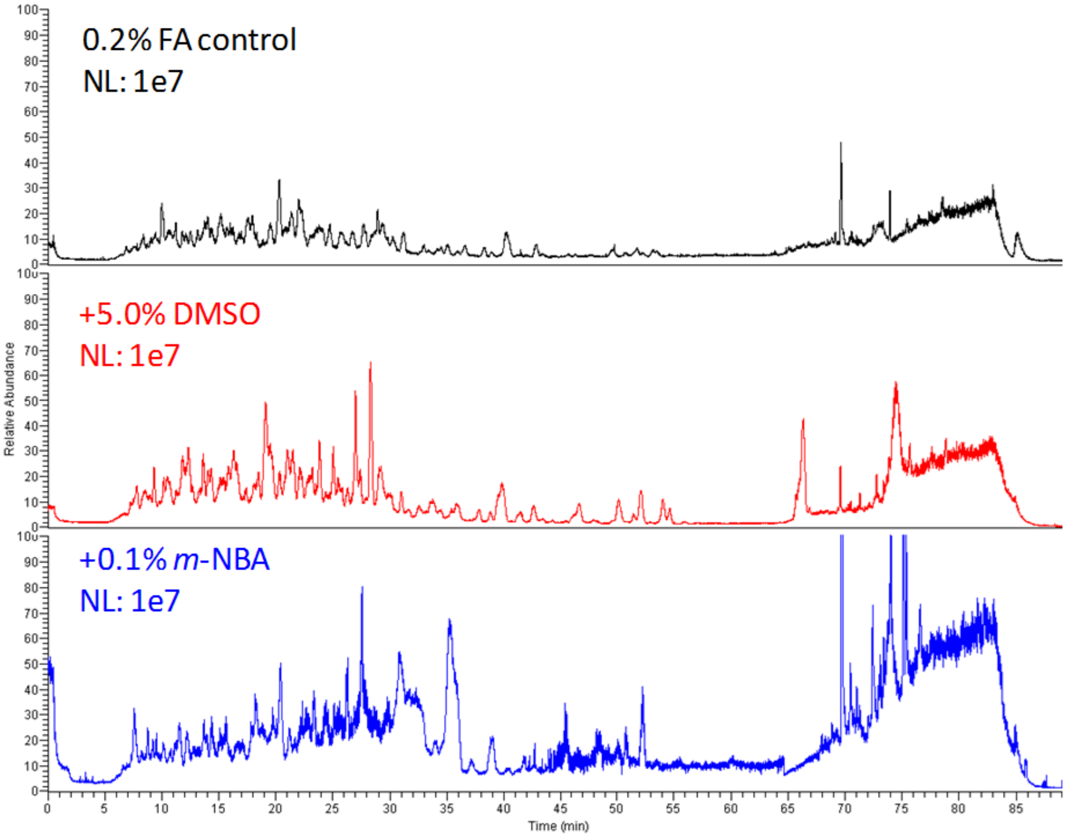


Figure 2.1. Comparison of TICs observed for each mobile phase condition. Traces are from full-scan-only experiments of the 5 protein mixture digested with pepsin (16 ng total protein injected). The chromatograms were normalized so that for each of them, 100% on the y axis equals 10^7 total ions.

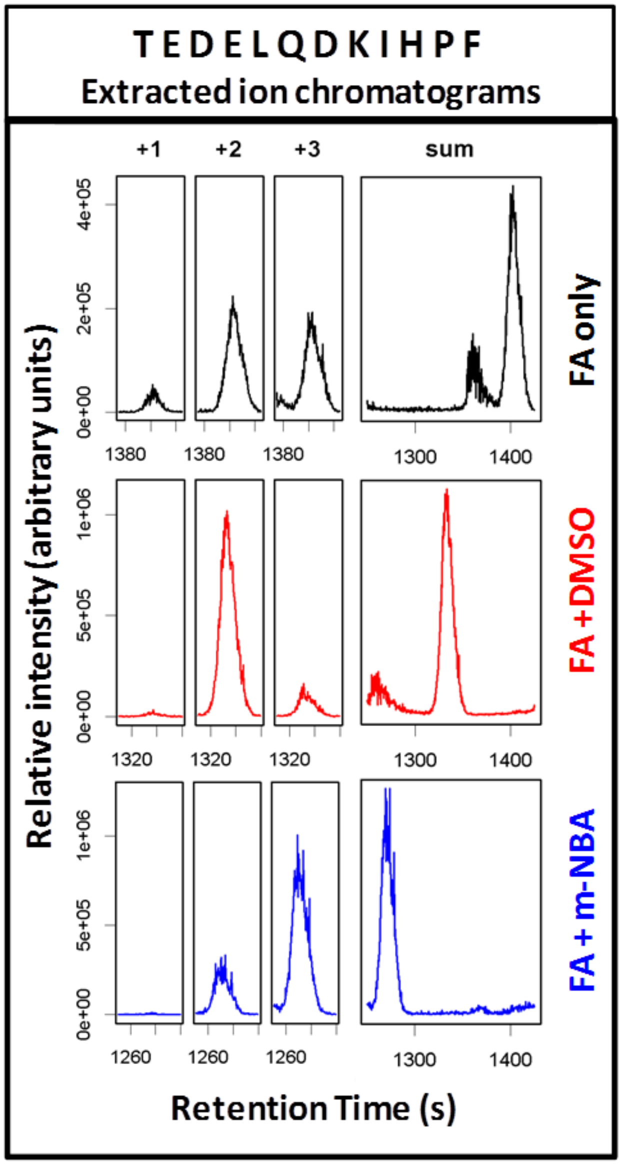


Figure 2.2. Extracted ion chromatograms (EICs) for the +1, +2, and +3 charge states of the elastic peptide TEDELQDKIHPF, illustrating the relative charge distributions produced during ESI for three mobile phases: 0.2% FA only (top, black), 0.2% FA + 5.0% DMSO (middle, red), or with 0.2% FA + 0.1% *m*-NBA (bottom, blue).

Table 2.2. Comparison of chromatographic FWHM for peptides produced by each protease analyzed in the three different solvent systems.

	TEDELQDKIHPF	TYFPHFDSHGSAQVK	SDIPNPIGSENS
0.2% FA only	14.1 s	17.6 s	19.4 s
+ 5.0% DMSO	13.6 s	21.5 s	19.8 s
+ 0.1% m-NBA	15.6 s	22.7 s	25.6 s

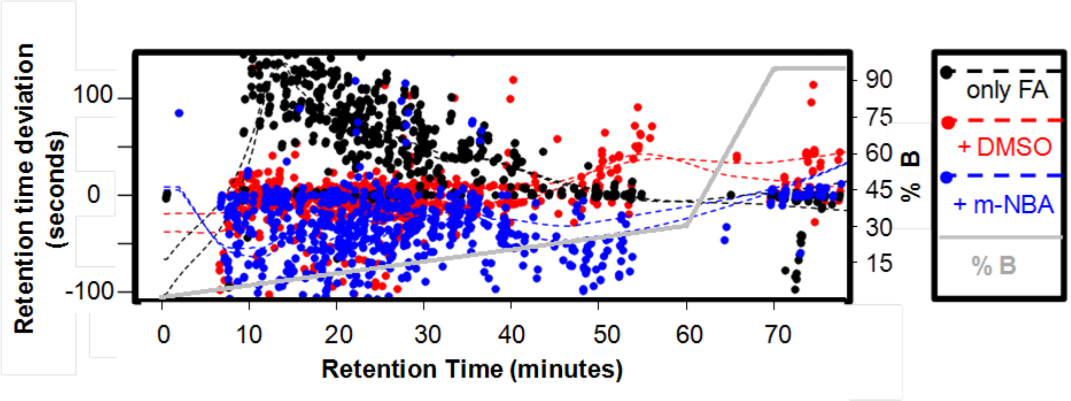


Figure 2.3. Global effects of supercharging reagents DMSO and *m*-NBA on chromatographic peptide retention during reversed-phase nano LC-ESI-MS/MS. Peaks identified in all samples were aligned and the median observed retention was chosen as the zero point. Both DMSO and *m*-NBA modified mobile phases generally result in reduced retention times as compared to the control, 0.2% FA alone (black). At high concentrations of ACN, the DMSO caused increased retention (red), whereas the *m*-NBA continued to reduce retention times (blue). The solid grey line gives the gradient profile, and the dashed lines represent the moving average of retention time deviation for each condition.

NBA generally decreased retention times compared to the FA only control. m-NBA introduces different functional groups to the reversed-phase separation system that allow hydrogen bonding, aromatic pi-stacking/pi-cation interactions, and ionic interactions with the nitro group. Together these properties appear to decrease the quality of chromatography in the presence of m-NBA.

3. Tandem mass spectrum quality

The quality of MS/MS spectra was assessed for both CID and ETD fragmentation of the doubly charged precursor ion and the triply charged precursor ion for the elastic peptide, TEDELQDKIHPF (Figure 2.4). As expected, a significant improvement in the ETD fragment ion series resulting from the triply charged precursor ion as compared to the doubly charged precursor ion was observed (Syka et al. 2004). The use of supplemental activation resulted in significant populations of b- and y-ions in the ETD fragment ion spectra (Swaney et al. 2006).

Interestingly, the triply charged precursor ion was poorly fragmented by CID, and the spectrum did not match the sequence below 1% FDR. Thus, for this example, the peptide identification was of high quality for the doubly charged precursor ion by CID and for the triply charged precursor ion by ETD, which highlights the complementary nature of the two approaches. To examine whether in each co-solvent more identifications were made by ETD or CID, we searched CID and ETD spectra separately (Table 2.3, Figure 2.5). MS-GFDB can either score CID and ETD spectra separately, or merge the CID and ETD spectra from the same precursor into a summed, scored spectrum (Kim et al. 2010). When the spectra were searched separately, DMSO afforded increased numbers of peptides identified in both CID and

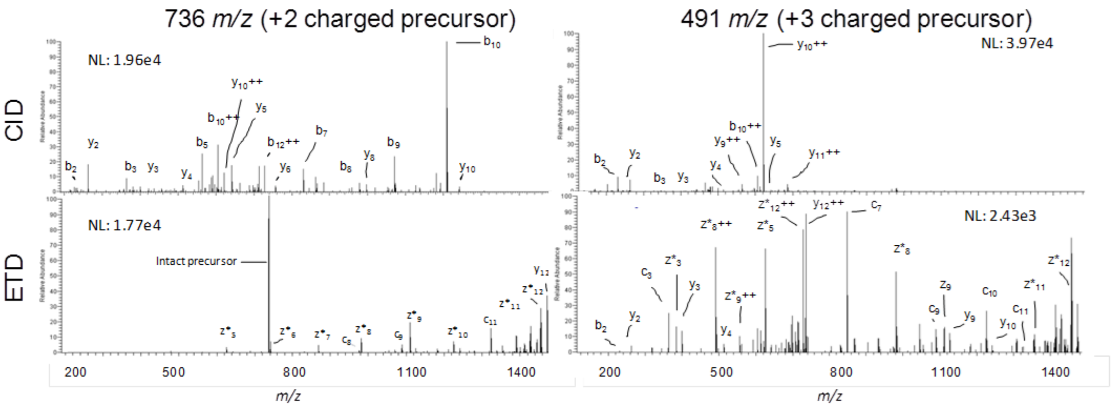


Figure 2.4. MS/MS spectra produced from the fragmentation of the doubly and triply charged precursor for the elastic peptide TEDELQDKIHPF. CID of the doubly charged precursor produced a rich fragment ion spectra, but CID of the triply charged precursor resulted in few fragment ions. In contrast, ETD of the doubly charged precursor produced a weak ladder of fragment ions, but ETD of the triply charged precursor produced a nearly complete sequence ladder of *c*- and *z**- ion pairs. Supplemental activation used with ETD also resulted in *b*-, *y*- ions that aided in sequence determination.

Table 2.3. Unique peptide counts for each mobile phase divided by activation.

Protease	activation	0.2% FA	FA + DMSO	FA + mNBA
Trypsin	CID	164 ± 8	201 ± 7	142 ± 8
	ETD	165 ± 3	190 ± 2	153 ± 10
	Total unique	178 ± 5	219 ± 2	165 ± 8
Elastase	CID	90 ± 5	103 ± 7	62 ± 13
	ETD	111 ± 6	121 ± 10	90 ± 16
	Total unique	121 ± 7	136 ± 9	100 ± 21
Pepsin	CID	120 ± 20	132 ± 10	88 ± 15
	ETD	177 ± 25	190 ± 22	163 ± 28
	Total unique	200 ± 30	221 ± 24	178 ± 34

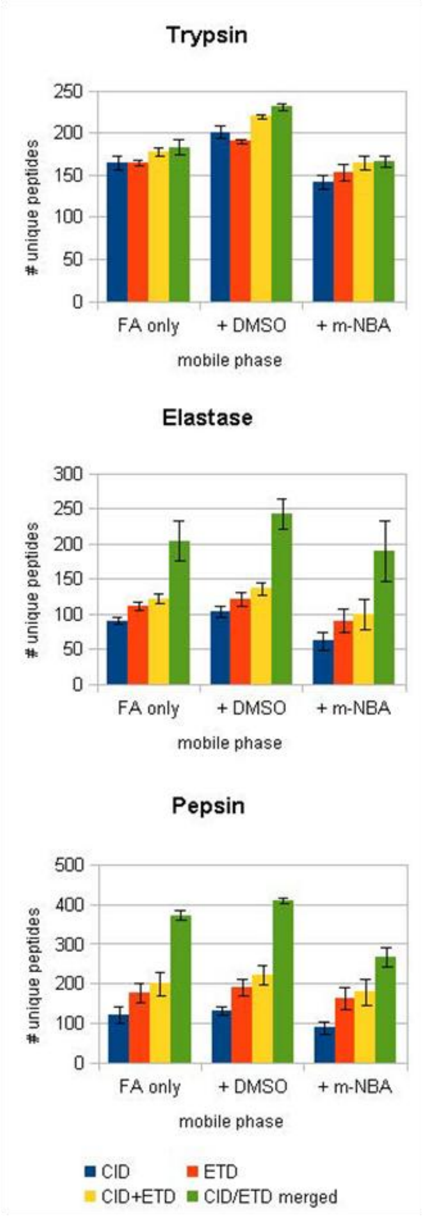


Figure 2.5. Comparison of unique peptide counts for each protease and each mobile phase condition. The number of unique peptides identified by CID (blue) is compared to the number of unique peptides identified by ETD (red). The sum of unique peptides identified from both CID and ETD spectra combined after the database search (yellow) is also compared to the number of unique peptides identified using the option to merge spectra from the same precursor before computing the score histogram (green). For non-tryptic peptides, the merged scoring produced dramatic increases in the number of unique peptides that are identified. The error bars are +/- one standard deviation of the average of three independent experiments.

ETD whereas fewer peptides were identified from the m-NBA co-solvent as compared to FA alone. Similar results were obtained from the searches in which the CID and ETD spectra were merged. In addition, dramatic gains in identifications were achieved using the merged search for non-tryptic peptides.

4. Effects on peptide charge state

The effects of DMSO and m-NBA on peptide precursor charge state distributions were assessed with data from full-scan-only experiments. Extracted ion chromatograms (EICs) for all possible charge states of peptides identified from MS/MS experiments were analyzed. Representative EICs for the singly, doubly, and triply charged precursor ions of an elastic peptide (TEDELQDKIHPF) are shown in Figure 2.2. Peaks from EICs were integrated and areas for each charge state were used to calculate the charge distribution for four peptides containing zero, one, two, or three basic side-chains (Table 2.4). For 0.2% FA alone, the average charge state was centered near the number of basic functional groups present within the peptide, but a significant portion of the peptide also was found with one more charge. When m-NBA was added, most of the peptide carried the additional charge. With DMSO as the co-solvent, the charge state distribution coalesced to the charge state corresponding to the number of basic residues. Results from experiments on intact proteins also showed decreases in charge state relative to control at low concentrations of DMSO (Sterling et al. 2011).

To globally assess the effects of the charge state distributions produced using each mobile phase condition on overall quality of data produced, we assessed the

Table 2.4. Peptide charge state distributions for peptides with various positive side chain counts.

		Peptide 0	Peptide 1	Peptide 2	Peptide 3
0.2% FA only	+1	68	08	07	01
	+2	32	92	51	21
	+3	n.d.*	00	42	39
	+4	n.d.*	n.d.*	n.d.*	39
summed intensity		1.0e6	8.4e6	5.6e6	2.0e7
+ 5.0% DMSO	+1	97	02	00	00
	+2	03	98	90	14
	+3	n.d.*	00	10	70
	+4	n.d.*	n.d.*	n.d.*	16
Summed intensity		1.6e6	1.7e7	1.5e7	4.2e7
+ 0.1% m-NBA	+1	04	00	00	00
	+2	96	60	21	00
	+3	n.d.*	40	78	25
	+4	n.d.*	n.d.*	n.d.*	75
Summed intensity		2.0e6	1.7e7	1.4e7	4.4e7

*n.d. = not detected. Normalized ratio of integrated peak area from the EIC of each charge state of each peptide, numbered in order of potential positive charge sites. Each value is the average of two replicate injections. Peptide 0: Peptic peptide containing no positive charge-bearing side-chains, SDIPNPIGSENS. Peptide 1: Tryptic peptide containing only one positive residue, YNGVFQECCQAEDK. Peptide 2: Elastase generated peptide containing two positive residues, TEDELQDKIHPF. Peptide 3: Tryptic peptide containing one lysine and two histidines positive side-chains, TYFPHFDSHGSAQVK.

number of unique peptides identified and total spectral counts for each peptide at each charge state from each mobile phase condition (Table 2.5). Both DMSO and m-NBA co-solvents resulted in nearly double the total ion signal (as calculated from EICs generated for all charge states summed). However, addition of DMSO resulted in a significant increase in the number of peptides identified. This increase appeared to be due to a significant signal enhancement for the most probable charge state. Thus, DMSO causes charge state coalescence into a single charge state, which translates into simpler precursor spectra resulting in improved MS/MS data.

For all peptides, the m-NBA modified mobile phase resulted in the highest precursor charge states. m-NBA was able to supercharge peptides to charge states greater than the number of basic functional groups present (i.e. R, K, H, and the n-terminal). For example, ESI of Glu-fib, EGVNDNEEGFFSAR, resulted in almost exclusively a doubly charged precursor ion for 5.0% DMSO and the FA only control, but with 0.1% m-NBA, the precursor ion was mostly triply charged. This peptide has only one basic residue and four acidic residues. The basic residue and the N-terminus are expected to carry positive charges; however, the third site of charging in the m-NBA co-solvent is not obvious. The utility of m-NBA supercharging is further demonstrated by the identification of a peptic peptide that did not contain any positive side chains (peptide 0 in Table 2), which was only sequenced in the analysis carried out with m-NBA added (Figure 2.6). In fact, 17 peptic peptides that did not contain R, K, or H residues were identified using m-NBA as the co-solvent. Therefore, supercharging, at least with peptide analytes, can cause peptide precursor ions to have a number of charges greater than the number of basic functional groups present. This

Table 2.5. Unique peptides identified and spectral counts according to charge state for each protease

Trypsin	only FA		FA + DMSO		FA + m-NBA	
Z	unique pep.	Sp. cts.	unique pep.	Sp.cts.	unique pep.	Sp. cts.
2	121 ± 8	347 ± 24	162 ± 10	492 ± 24	67 ± 4	167 ± 2
3	57 ± 3	194 ± 8	61 ± 7	161 ± 16	86 ± 6	211 ± 13
4	4 ± 1	36 ± 3	8 ± 2	31 ± 2	11 ± 3	53 ± 2
5	0 ± 1	8 ± 2	0 ± 0	4 ± 3	2 ± 0	6 ± 1
Elastase	only FA		FA + DMSO		FA + m-NBA	
Z	unique pep.	Sp. cts.	unique pep.	Sp. cts.	unique pep.	Sp. cts.
2	106 ± 21	289 ± 59	135 ± 11	320 ± 34	85 ± 28	201 ± 71
3	89 ± 7	273 ± 12	102 ± 10	321 ± 47	78 ± 12	216 ± 46
4	9 ± 2	32 ± 4	5 ± 2	28 ± 8	25 ± 4	62 ± 13
5	0 ± 1	0 ± 1	1 ± 1	1 ± 1	0 ± 1	0 ± 1
Pepsin	only FA		FA + DMSO		FA + m-NBA	
Z	unique pep.	Sp. cts.	unique pep.	Sp. cts.	unique pep.	Sp. cts.
2	158 ± 12	319 ± 15	175 ± 8	413 ± 20	88 ± 5	205 ± 17
3	130 ± 5	280 ± 5	136 ± 54	319 ± 15	106 ± 13	256 ± 24
4	80 ± 5	200 ± 8	87 ± 1	182 ± 7	45 ± 5	120 ± 24
5	4 ± 1	16 ± 2	9 ± 2	17 ± 1	26 ± 5	65 ± 12

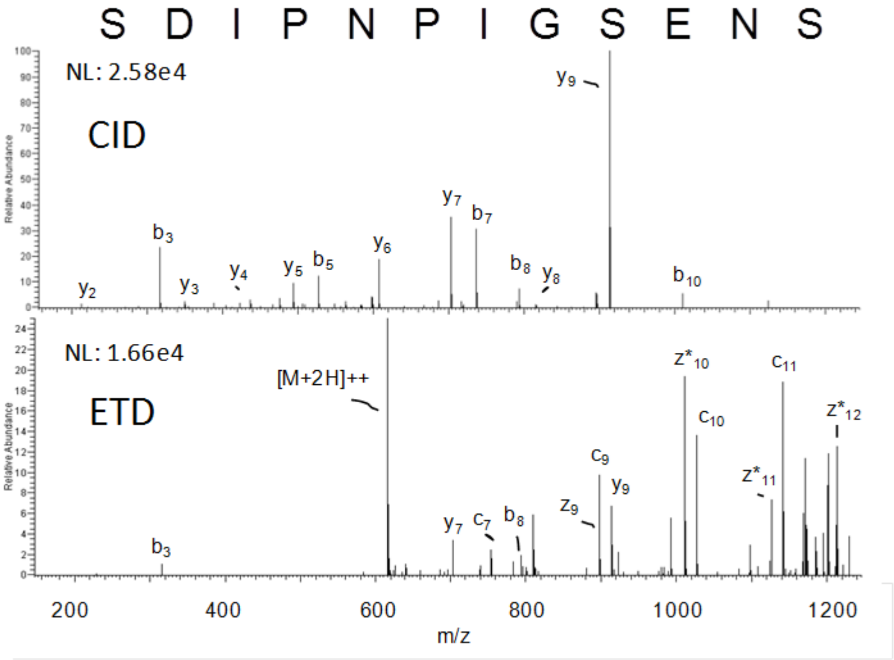


Figure 2.6. MS/MS spectra produced from the fragmentation of the doubly charged precursor of the peptic peptide: SDIPNPIGSENS, which bears no positive charged side chains. (A) CID fragmentation results in a spectra dominated by the y_9 ion, which corresponds to fragmentation at proline. (B) ETD with supplemental activation results in fragments that complement the CID spectra in the high mass region.

observation is in contrast to speculation by Douglass and Venter that the number of basic residues may limit the extent of supercharging (Douglass and Venter 2012).

To assess the potential benefit from m-NBA-mediated charge enhancement, a theoretical digest was carried out for each of the proteases on the five protein mixture using MS-digest in Protein Prospector (www.prospector.ucsf.edu). Peptides produced from trypsin digestion can only lack a positive side chain if they arise from the protein C-terminus. In fact, only 1% of theoretical tryptic peptides from the five protein mixture lack any basic amino acid side chains. Elastase and pepsin, however, are predicted to result in more peptides lacking possible positive charges, 14% and 12%, respectively.

4. Conclusions

Here we present an analysis of factors relevant to the application of supercharging reagents DMSO and m-NBA for peptide identification by nano-ESI-LC-MS/MS. Consistent with previous reports, the m-NBA modified mobile phase produced extensive supercharging of peptides. However, the extent of charge enhancement did not correlate with the number of peptide identifications even with sequential fragmentation with both CID and ETD. The addition of m-NBA was very important for obtaining better ETD spectra from peptides that did not have many side chains that could carry a positive charge, and therefore, may find use in targeted experiments. For data-dependent, untargeted experiments however, the lower chromatographic quality and the broad distribution of highly charged precursors achieved with the m-NBA co-solvent obviated any improvements in total numbers of high quality peptide identifications. Kjeldsen et al also observed chromatographic broadening but did not observe a decrease in the number of identifications from a digest of BSA (Kjeldsen et al. 2007). Chromatographic broadening is expected to adversely affect analyses of more complex mixtures but may not affect the analysis of simple ones. In addition, the broad precursor charge state distribution results in several precursor ions for each peptide in the complex mixture, resulting in a lower signal to noise ratio. Further, highly charged precursors result in multiply charged fragment ions that are difficult to resolve on low resolution ion-trap instruments.

Using a combination of ion mobility and circular dichroism, Sterling et al. showed recently that supercharging reagents act as chemical denaturants for intact proteins, and that the higher charge species are more unfolded (Sterling et al. 2010). In the case of peptides, it is likely that they are completely unfolded and that the charge state will depend on the number of basic residues as well as the length of the peptide. Therefore, we postulate that the number of peptide microstates allowed in the electrosprayed droplet containing DMSO is limited to only the most favorable extent of charging allowed by each individual peptide sequence and length. Thus, although supercharging was observed with the m-NBA, it was not observed with DMSO.

It is interesting to speculate on the reasons for the marked charge coalescence with DMSO. Visual comparison of the electrospray under each tested mobile phase condition showed that the most stable spray was achieved across all parts of the gradient with DMSO-modified mobile phases. According to the Raleigh equation, spray stability could also contribute to more uniform droplet sizes which might translate directly into more uniform charge states. Finally, it is possible that DMSO acts to promote desolvation efficiency and increase the total signal at a favorable charge state.

For both tryptic and non-tryptic peptides, DMSO increases total MS/MS productivity apparently due to charge state coalescence, which results in more peptide precursor signal at a predominant charge state. DMSO is expected to be particularly useful when complex peptide mixtures are analyzed using data-dependent acquisition approaches. Therefore, we expect this simple mobile phase addition to be widely adopted in bottom-up peptide identification experiments, regardless of protease and fragmentation. Further gains are expected when this strategy is combined with high resolution mass spectrometry.

Chapter II, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in Journal of the American Society for Mass Spectrometry. (Meyer, J.G. and Komives, E. A. (2012). Charge State Coalescence During Electrospray Ionization Improves Peptide Identification by Tandem Mass Spectrometry. Journal of the American Society for Mass Spectrometry, 23, 1390-1399.)

D. References

Ahadi, E. and L. Konermann (2011). "Modeling the Behavior of Coarse-Grained Polymer Chains in Charged Water Droplets: Implications for the Mechanism of Electrospray Ionization." J Phys Chem B.

Burkhart, J. M., C. Schumbrutzki, S. Wortelkamp, A. Sickmann and R. P. Zahedi (2011). "Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics." J Proteomics **in press**.

Coulier, L., R. Bas, S. Jespersen, E. Verheij, M. J. van der Werf and T. Hankemeier (2006). "Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography−Electrospray Ionization Mass Spectrometry." Anal Chem **78**(18): 6573-6582.

de Godoy, L. M. F., J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther and M. Mann (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature **455**(7217): 1251-1254.

Douglass, K. A. and A. R. Venter (2012). "Investigating the role of adducts in protein supercharging with sulfolane." J Am Soc Mass Spectrom **23**(3): 489-497.

Fenn, J., M. Mann, C. Meng, S. Wong and C. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules " Science **246**(4926): 64 -71.

Ficarro, S. B., Y. Zhang, Y. Lu, A. R. Moghimi, M. Askenazi, E. Hyatt, E. D. Smith, L. Boyer, T. M. Schlaeger, C. J. Luckey and J. A. Marto (2009). "Improved Electrospray Ionization Efficiency Compensates for Diminished Chromatographic Resolution and Enables Proteomics Analysis of Tyrosine Signaling in Embryonic Stem Cells." Anal Chem **81**(9): 3440-3447.

Gustavsson, S. Å., J. Samskog, K. E. Markides and B. Långström (2001). "Studies of signal suppression in liquid chromatography–electrospray ionization mass spectrometry using volatile ion-pairing reagents." J Chromatog A **937**(1-2): 41-47.

Hogan, J. C. J. and P. Biswas (2008). "Monte Carlo Simulation of Macromolecular Ionization by Nanoelectrospray." J Am Soc Mass Spectrom **19**(8): 1098-1107.

Iavarone, A. T. and E. R. Williams (2002). "Supercharging in electrospray ionization: effects on signal and charge." Int J Mass Spectrom **219**(1): 63-72.

Iavarone, A. T. and E. R. Williams (2003). "Mechanism of Charging and Supercharging Molecules in Electrospray Ionization. ." J Am Chem Soc **125**(8): 2319-2327.

Kamel, A. M., P. R. Brown and B. Munson (1999). "Effects of Mobile-Phase Additives, Solution pH, Ionization Constant, and Analyte Concentration on the Sensitivities and Electrospray Ionization Mass Spectra of Nucleoside Antiviral Agents." Anal Chem **71**(24): 5481-5492.

Keller, A., J. Eng, N. Zhang, X.-j. Li and R. Aebersold (2005). "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." Mol Syst Biol **1**(2005.0017).

Kessner, D., M. Chambers, R. Burke, D. Agus and P. Mallick (2008). "ProteoWizard: open source software for rapid proteomics tools development." Bioinformatics **24**(21): 2534-2536.

Kim, M.-S., J. Zhong, K. Kandasamy, B. Delanghe and A. Pandey (2011). "Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides." Proteomics **11**(12): 2568-2572.

Kim, S., N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck and P. A. Pevzner (2010). "The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search." Mol Cell Proteom **9**(12): 2840-2852.

Kjeldsen, F., A. M. Giessing, C. R. Ingrell and O. N. Jensen (2007). "Peptide sequencing and characterization of post-translational modifications by enhanced ion-charging and liquid chromatography electron-transfer dissociation tandem mass spectrometry." Anal Chem **79**(24): 9243-9252.

Lin, Y., J. Zhou, D. Bi, P. Chen, X. Wang and S. Liang (2008). "Sodium-deoxycholate-assisted tryptic digestion and identification of proteolytically resistant proteins." Anal Biochem **377**(2): 259-266.

Lomeli, S. H., I. X. Peng, S. Yin, R. R. Ogorzalek Loo and J. A. Loo (2010). "New Reagents for Increasing ESI Multiple Charging of Proteins and Protein Complexes." J Am Soc Mass Spectrom **21**(1): 127-131.

Loo, J. A., H. R. Udseth, R. D. Smith and J. H. Futrell (1988). " Collisional effects on the charge distribution of ions from large molecules, formed by electrospray-ionization mass spectrometry. ." Rapid Commun Mass Spectrom **2**(10): 207-210.

Loo, R. R. O. and R. D. Smith (1995). "Proton transfer reactions of multiply charged peptide and protein cations and anions." J Mass Spectrom **30**(2): 339-347.

Mallet, C. R., Z. Lu and J. R. Mazzeo (2004). "A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts." Rapid Commun Mass Spectrom **18**(1): 49-58.

Motoyama, A. and J. R. Y. III (2008). "Multidimensional LC Separations in Shotgun Proteomics." Anal Chem **80**(19): 7187-7193.

Ong, S.-E. and M. Mann (2005). "Mass spectrometry-based proteomics turns quantitative." Nat Chem Biol **1**(5): 252-262.

Rayleigh, L. (1882). Phils Mag **14**: 184-186.

Sleno, L. and D. A. Volmer (2004). "Ion activation methods for tandem mass spectrometry." J Mass Spectrom **39**(10): 1091-1112.

Smith, C. A., E. J. Want, G. C. Tong, R. Abagyan and G. Siuzdak (2006). "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. ." Anal Chem **78**(3): 779-787.

Sterling, H. J., M. P. Daly, G. K. Feld, K. L. Thoren, A. F. Kintzer, B. A. Krantz and E. R. Williams (2010). "Effects of Supercharging Reagents on Noncovalent Complex Structure in Electrospray Ionization from Aqueous Solutions." J Am Soc Mass Spectrom **21**(10): 1762-1774.

Sterling, H. J., J. S. Prell, C. A. Cassou and E. R. Williams (2011). "Protein conformation and supercharging with DMSO from aqueous solution." J Am Soc Mass Spectrom **22**(7): 1178-1186.

Swaney, D. L., G. C. McAlister and J. J. Coon (2008). "Decision tree-driven tandem mass spectrometry for shotgun proteomics." Nat Meth **5**(11): 959-964.

Swaney, D. L., G. C. McAlister, M. Wirtala, J. C. Schwartz, J. E. P. Syka and J. J. Coon (2006). "Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors." Anal Chem **79**(2): 477-485.

Swaney, D. L., C. D. Wenger and J. J. Coon (2010). "Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics." J Proteom Res **9**(3): 1323-1329.

Syka, J. E. P., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." Proc Nat Acad Sci USA **101**(26): 9528 -9533.

Tabb, D. L., L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin and J. R. Yates (2003). "Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides." Anal Chem **75**(5): 1155-1163.

Team, R. D. C. (2011). R: A Language and Environment for Statistical Computing.

Valeja, S. G., J. D. Tipton, M. R. Emmett and A. G. Marshall (2010). "New Reagents for Enhanced Liquid Chromatographic Separation and Charging of Intact Protein Ions for Electrospray Ionization Mass Spectrometry." Anal Chem **82**(17): 7515-7519.

Wilm, M. (2011). "Principles of Electrospray Ionization." Mol Cell Proteom **10**(7): 1-8.

Yin, S. and J. A. Loo (2011). "Top-down mass spectrometry of supercharged native protein–ligand complexes." Int J Mass Spectrom **300**(2-3): 118-122.

Chapter III

Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases

A. Introduction

The most powerful technique for system-scale protein measurement, or proteomics, is mass-spectrometry based proteomics (Walther and Mann 2010). Although great progress has enabled quantification of nearly all proteins expressed in yeast (de Godoy et al. 2008; Nagaraj et al. 2012), sequence coverage is often dismal with some proteins being identified by a single peptide sequence. Complete amino acid coverage is valuable for comprehensive profiling of post-translational modifications (e.g. phosphorylation) and for quantification of splice variants. Low observed proteome coverage is due to several factors including the wide dynamic range of protein concentrations in biological samples, splice variants, and unanticipated or unconsidered post translations modifications (PTMs). Improvements to every step of the bottom-up proteomics workflow continue to increase the observable proteome.

Due to length constraints that limit observable peptides, proteome coverage is ultimately limited by the proteome digestion. Typically, identifiable peptides are between 7 and 35 amino acids with the lower limit being determined by sequence uniqueness, and the upper limit being determined by instrument resolving power (Swaney et al. 2010). *In silico* proteome digestions predict that nearly one quarter of peptides generated from tryptic digestion of the S. cerevisiae proteome will be only a single amino acid long. Sequences lost due to length overall result in a theoretical upper proteome coverage limit of 68.8% according to in silico predictions (Figure3.1).

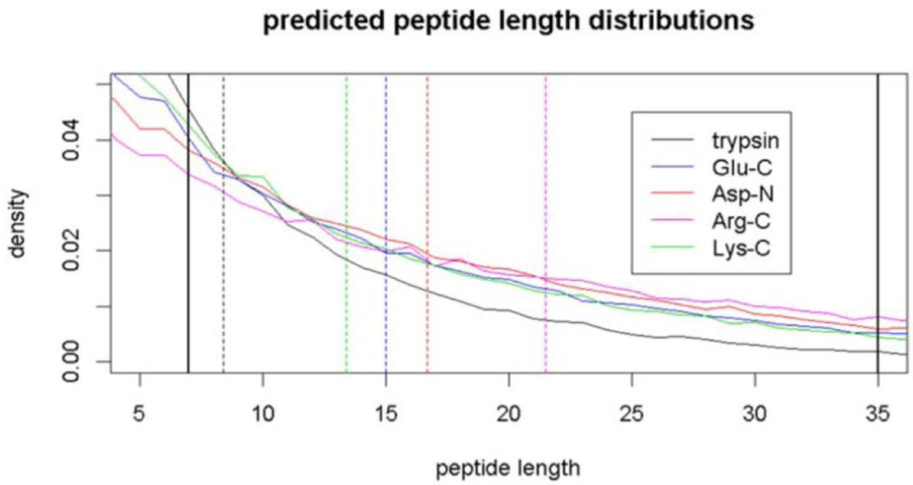


Figure 3.l. Predicted length distributions of peptides generated by in silico digestion of the S. cerevisiae proteome with trypsin, Glu-C, Asp-N, Arg-C, and Lys-C.

Recently, several groups have demonstrated that combining data from separate protease digestions improves proteome coverage (MacCoss et al. 2002; Wang et al. 2008; Rietschel et al. 2009; Swaney et al. 2010). Improved peptide yield was also shown allowing proteome analysis of small-quantity samples from laser-capture microdissection (Wiśniewski et al. 2012; Wiśniewski and Mann 2012). Swaney, Wenger and Coon used trypsin, Lys-C, Arg-C, Glu-C, and Asp-N to double the observed S. cerevisiae non-redundant amino acid coverage from 11.9% to 25.5% (Swaney et al. 2010).

Other proteases that are used in proteomics to complement trypsin mainly cleave at ionic amino acid side chains and it would be useful to have proteases with additional, complementary specificities. Here we demonstrate the application of wild-type alpha-lytic protease (WaLP) (Silen et al. 1989), and an active site mutant of WaLP, M190A alpha-lytic protease (MaLP) (Bone et al. 1989), to proteome digestion for shotgun proteomics. Both were reported to have specificity for cleaving after aliphatic side chains, which are more common amino acids. WaLP is a serine protease secreted from the soil bacteria Lysobacter enzymogenesis (Silen et al. 1989; Graham et al. 1993) and has been studied extensively by mutagenesis and biophysical methods (Bone et al. 1989). WaLP was found to exhibit remarkable stability (Sohl et al. 1998; Jaswal et al. 2002).

Non-tryptic peptides are more difficult to identify than tryptic peptides, especially when lacking defined termini (i.e. from semi-specific protease digestion or endogenous peptides) due to increased database search space and less predictable ionization and fragmentation. Lack of defined termini drastically increases database search space because more possible peptides fall within the precursor tolerance and drive up false positive rates (Gupta et al. 2011). The majority of tryptic peptides have one positive charge localized at each termini upon +2 precursor charge state upon electrospray ionization (ESI), which results in well characterized fragmentation by collision induced dissociation (CID) (Wysocki et al. 2000; Tabb et al. 2003). Non-tryptic peptides, in contrast, may lack positively charged side chains (i.e. Arg, Lys, His) altogether, thereby making it unlikely to obtain multiple charges upon ESI. Those that do contain positive charges away from the C-terminus produce less predictable fragmentation upon CID. Recently, additional peptide fragmentation methods have become accessible, such as electron-transfer dissociation (ETD) (Syka et al. 2004), which produces fragment ion series that are less dependent on peptide sequence, and higher-energy collisional dissociation (HCD) (Michalski et al. 2012). An in-depth comparison of activation methods for non-tryptic peptide identification has been published recently by Smith's lab, where the authors evaluated FT-CID, and FT- ETD, and FT-HCD for sequencing peptides isolated from blood plasma (Shen et al. 2011).

To enable application of the alpha-lytic proteases which have specificity for aliphatic amino acid side chains to shotgun proteomics, we address the above issues by comparing multiple fragmentation modes in combination with the peptide identification algorithm, MS-GFDB, which easily learns scoring parameters from an initial set of annotated peptide-spectra matches (PSMs) for arbitrary fragmentation methods and proteases (Kim et al. 2010). We analyzed standard protein mixtures and complex *S. pombe* proteomes digested with trypsin, LysC, WaLP, and MaLP.

Specifically, we assessed ion activation methods, observed peptide character, and biological gains additional due to additional digestions. The results present the pros and cons of using orthogonal proteases in proteomics.

B. Materials and Methods

1. Samples and Chemicals

Acetonitrile (ACN) optima, tri-carboxyethyl phosphine (TCEP, Bond-breaker), and HPLC-grade ethyl-acetate were purchased from Thermo-Fisher Scientific (Waltham, MA). N-ethyl maleimide (NEM), formic acid (FA), sodium deoxycholate (SDC), sodium dodecyl sulfate (SDS), and Trizma-brand tris buffer were from Sigma-Adrich (St. Louis, MO). All chemicals were the highest grade available and were used without further purification. Sequencing-grade modified trypsin was purchased from Promega (Madison, WI). A mixture of standard proteins was prepared containing bovine aprotinin (6.5 kDa, P00974), murine leptin (16 kDa, P41160), horseradish peroxidase (39 kDa, P00433), *E. coli* GroEL (57 kDa, P0A6F5), bovine serum albumin (69 kDa, P02769), cytochrome c (12.4 kDa, P00004), and hemoglobin α 15.3 kD, P69905) and β chains (16 kDa, P68871). This mixture was prepared as described and compared with recently published results (Guthals et al. 2013).

2. Protease expression and purification

WaLP was expressed from *Lysobacter enzymogenesis* type 495 using Bachovichin's media supplemented with MEM vitamins and 60g/L sucrose. *L.*

*enz.*was grown at 30 °C with shaking at 100 rpm for 3 days. MaLP was expressed as described previously (Mace et al. 1995) in D1210 E. coli using the pALP12-ΔM190A plasmid, which was the generous gift of Dr. Dave Agard. Both proteases were purified from the culture supernatant as described previously (Mace and Agard 1995). Briefly, the protease is captured from the supernatant by batch binding on SP-sepharose, which is washed extensively and then eluted with high pH glycine buffer. After buffer-exchange to pH 7.2, the enzyme was loaded by superloop onto the FPLC monoS column using a gradient of 10 mM $NaHPO_4$, pH 7.2 to the same buffer containing 250 mM sodium acetate over 1 hr.

3. Protease activity assays

Enzyme activity was assessed with a chromogenic assay using N-succinyl-Ala-Ala-Ala-p-nitroanalide (Sigma-Adrich, St. Louis, MO)) for WaLP or N-Succinyl-Ala-Ala-Pro-Leu-p-nitroanilide (Bachem Americas, Torrance, CA) for MaLP or N-Succinyl-Ala-Ala-Pro-Phe-p-nitroanilide (Sigma-Adrich, St. Louis, MO) for chymotrypsin or N-p-tosyl-Gly-Pro-Lys-p-nitroanilide (Sigma-Adrich, St. Louis, MO) for trypsin and LysC. The specific activity of WaLP was $5x10^{-4}$ mmols N-succinyl-Ala-Ala-Ala-p-nitroanalide hydrolyzed per min per mg WaLP, and the specific activity of MaLP was $3x10^{-2}$ mmols N-Succinyl-Ala-Ala-Pro-Phe-p-nitroanilide hydrolyzed per min per mg MaLP. The WaLP and MaLP proteases are both being made available from Sigma Aldrich. All protease assays were carried out under identical buffer conditions, except that SDC assays were done in HEPES because tris-buffered SDC without dissolved protein is very viscous.

4. In gel digestion

To test the suitability of WaLP and MaLP for in-gel digests, we obtained a sample of glucose transporter-5 (Uniprot Acc# P22732) that was expressed in *Pichia pastoris* and then deglycosylated with PNGase F. After SDS PAGE, the band was excised and subjected to in-gel digestion separately with either trypsin, WaLP, or MaLP according to standard protocol (available at massspec.ucsd.edu/bioms/training/protocols.php). Resulting peptides were analyzed with the 5600 TripleTof (ABSCIEX) interfaced with a Waters NanoAcquity UPLC. Peptides were separated with a 1 hour, linear gradient from 5 to 80% mobile phase B at a flow rate of 250 µl/min using a charged-surface hybrid C18 column (75 micron ID X 20 cm length, 2.5-µm particles, Waters). Mobile phase A was 98% water, 2% ACN, 0.1% formic acid, and 0.005% TFA and mobile phase B was100% ACN, 0.1% formic acid, and 0.005% TFA. Precursor spectra (400-1250 m/z) were collected for 0.25 s followed by MS/MS (50-2,000 m/z) of up to 50 of the most intense charge +2, +3, and +4 precursors for 2.4 s. The minimum intensity for MS/MS selection was 150 counts. Precursors were dynamically excluded for 4 seconds. The data were analyzed with Protein Prospector as described below.

5. Proteome preparation and digestion

S. pombe cell lysates were a generous gift from Dr. Paul Russell. *S. pombe* cells were lysed using a bead mill in 50mM Tris-HCl pH: 8.0; 150mM NaCl; 5mM EDTA; 10% Glycerol; 50mM NaF; 0.1mM Na3VO4; 0.2% NP40. Lysates were clarified at 15,700 x g for 10 min and the supernatant was removed. Insoluble material from the lysate was re-extracted according to a non-SDS compatible protocol, combined with the soluble material, and precipitated by chloroform/methanol extraction as described previously (von der Haar 2007). Protein precipitates were resuspended in 100mM Tris, pH 7.2 containing 1.0% sodium deoxycholate (SDC), reduced with 5 mM TCEP at 60 ^{o}C for 30 min, and alkylated with 10 mM N-ethylmaleimide (NEM) at room temperature for 1 hr. TCEP and NEM were then removed by ultrafiltration with a 10 kDa-cutoff amicon-4 (Millipore) with three 10-fold buffer exchanges into 100mM Tris, pH 7.2 containing 0.1% SDC. The alkylated *S. pombe* proteome concentration was determined using the BCA assay (Pierce Chemicals, from Thermo Scientific). Samples (150 µg) of *S. pombe* proteome were separately digested with either trypsin, LysC, WaLP, or MaLP at a ratio of 1:100 for 24 hrs at a total protein concentration of 0.5 mg/mL and SDC was removed by acidification with 5% formic acid (FA), extracted with ethylacetate, and purified by SepPak C18 (Waters, Inc) purification as described previously (Masuda et al. 2008; Masuda et al. 2009).

6. MS activation comparisons

A series of analyses of mixtures of known proteins and of unseparated proteome digests were performed in order to determine the best activation parameters for the MSMS runs. For these experiments, samples, 0.65 µg, of *S. pombe* proteome digest were resuspended in 5 µl 0.1% FA and injected onto a trap column (Waters' Symmetry 180 µm I.D. x 20 mm length, 5 µm C18 particles) equilibrated in 0.2% TFA, using a Waters NanoAcquity autosampler and binary solvent manager. A 100 µm I.D. X 15 cm column (packed in-house) containing 3 µm Magic C18 AQ particles was used for peptide separation using a 2.5 hour gradient of 2% to 30% B (0.2 % TFA in 90% ACN) at a flow rate of 0.6 µl/min. Total run time was 1.5 hour for the standard protein mix and 3 hour for the *S. pombe* digests including column flush and re-equilibration. Eluting peptides were electrosprayed at 2.7 kV using the Nanospray Flex Ion Source (Proxeon) into an LTQ-Orbitrap Velos hybrid mass spectrometer (ThermoFisher, Waltham, MA) using a precursor scan from 350-1400 m/z and a target resolution of 30,000 in profile mode. Unassigned and +1 precursor charge states were excluded, and dynamic exclusion was enabled for 45 sec allowing 1 repeat and using sequential activation of the top five precursors using CID, then ETD, then HCD with the FT mass analyzer. The scan rate for this experiment was 1.2 spectra/sec. Additional experiments were performed in which the top 10 precursors were sequentially targeted with CID then ETD using the ion trap as the mass analyzer as well as experiments in which the top 10 precursors were targeted using a data dependent decision tree (ddDT) approach (Swaney et al. 2008) to activate all +2 precursor charge states with CID and all +3 or greater precursor charge states with ETD. As expected, the faster scan rate of the ion trap yielded more peptide IDs compared to data from the higher resolution FT mass analyzer. The results from these experiments demonstrated the utility of the ddDT approach, which was then used to analyze the fully separated *S. pombe* proteome digests.

7. High-pH fractionation of proteome samples

Peptide fractionation by high-pH reverse-phase (HPRP) was performed as described previously (Wang et al. 2011). Briefly, lyophilized peptides were resuspended in 1.15mL of 20 mM NH_4HCO_2, pH 10 (HPRP mobile phase A). HPRP buffer B was 80% ACN with 20% 20 mM NH_4HCO_2, pH 10. Peptides were separated over a 100 X 2.1 mm Waters' C18 BEH column (5 µm particles) maintained at 40 ^oC.

Samples (1.05 ml) were loaded at a flow rate of 0.5 ml/min over 7 min in 98% A, and peptides were eluted with a gradient from 2% to 100% B over 27 min. Fifty-four 0.5 mL fractions were collected into 100 µl of 10% FA, and fractions were pooled according to the method of Smith's lab to yield 18 final pooled fractions that were lyophilized and stored at -80 ° C until nanoLC-MS/MS analysis (Wang et al. 2011).

8. NanoLC-ESI-MS/MS of HPRP-fractionated digests

Each pooled HPRP fraction was resuspended in 75 µl of 0.1% FA. Five µl (~0.5 µg/fraction) was injected into the LTQ-Orbitrap Velos hybrid mass spectrometer as described above except that a 60 min gradient from 2 to 30 percent B followed by column re-equilibration, for a total of 90 min per run. For these experiments, a ddDT (Swaney et al. 2008) was used to activate all +2 precursor charge states with CID and all +3 or greater precursor charge states with ETD. Total nLC-MS/MS acquisition time was 27 hours/protease, or 4.5 days total. Appendix table 3.1 contains a list of all of the experiments.

9. Database searches

Files (.RAW) were converted to .mzXML files using the default parameters msconvert.exe except for the option to centroid all spectra (version 3.0.4323, February 5th, 2013) within Trans-Proteomic Pipeline (TPP) (version 4.6.2) (Keller et al. 2005; Kessner et al. 2008). The standard protein mix data (CID/HCD/ETD triples, high resolution) was searched with Protein Prospector (prospector.ucsf.edu) against the E. coli subset of SwissProt (March 21st, 2012 version) with the sequences for each standard mix protein and protease added because the number of spectra was insufficient to properly train MS-GFDB. The database contained a total of 22,934 real and 22,934 randomized sequences comprising all E. coli strain sequences (45,868 total protein sequences) to allow estimation of the false discovery rate (FDR). Data from the unseparated *S. pombe* digests (CID/HCD/ETD triples, high resolution) was searched with Protein Prospector against the *S. pombe* subset of SwissProt (March 21st, 2012 version) with accessions for each protease added (4,990 real, 4,990 randomized, 10,980 total). An initial search was carried out with 10 ppm precursor tolerance and 15 ppm fragment-ion tolerance to calibrate the precursor masses, followed by another search with 5 ppm precursor tolerance and 15 ppm fragment-ion tolerance. Searches with trypsin and Lys-C data allowed up to 3 missed cleavages and one non-enzymatic termini. Searches of WaLP and MaLP data used “no enzyme” specificity. Default variable modifications were used. Searches required the fixed modification of cysteine with NEM. The data on unseparated *S. pombe* proteome collected as CID/HCD/ETD triples was also searched with MS-GFDB version 7780 (Kim et al. 2010) against common contaminants and the *S. pombe* complete proteome containing a total of 5099 real and 5099 reversed sequences (downloaded from UniProt on June 20th, 2012) using the merge search for comparison of the amount of internal ions. The comparison between Protein Prospector and MS-GFDB searches revealed that for WaLP and MaLP, Protein Prospector gave similar numbers of unique peptides (Appendix Table 3.2).

Data from the fully separated proteome analyses were converted to .mzXML and merged using mzXMLmerge (http://proteotools.pharmacy.arizona.edu/proteotools/index.jsp) to make database searching and downstream analysis more manageable. The merged .mzXML files were searched with MSGFplus.jar version 9352 (released on Feb, 4th 2013) (Kim et al., unpublished). MS-GFDB is a database search engine that reports rigorous p-values (spectral probabilities) for spectral interpretations based on all possible peptide match scores (Kim et al. 2010). The key advantages of the MS-GF algorithm are that it is highly effective in utilizing spectral evidence, the spectral interpretations are rigorously scored, and the scoring algorithm can be re-trained using large data sets of annotated spectra (Kim et al. 2008). MS-GFDB extends MS-GF to automatically derive scoring parameters from a set of annotated MS/MS spectra of any type (e.g. CID, ETD, etc.). This aspect was particularly important for efficient spectral interpretation of data from non-tryptic digests. MSGF+ is a successor of MS-GFDB that additionally allows input of mzml data and produces mzIdentML output files (Kim et al., unpublished). Database searches used default parameters except the number of tolerable enzymatic termini was set to 1 and searches of MaLP and WaLP used “no enzyme” specificity. Searches required fixed modification of cysteine by NEM and variable modification at peptide N-terminal Q to pyro-glutamate, protein N-terminal methionine loss plus acetylation, and methionine oxidation. Precursor masses containing between 0 and 2 ^{13}C were considered. For all MS-GFDB searches and all MS-GF+ searches, precursor mass tolerance was set to 5 ppm. After initial searches of each activation method alone, the scoring parameters were trained and the data were re-searched with the new scoring model. Only the MS-GFDB search engine was used for the large data sets because it was faster than Protein Prospector.

In order to quantitatively compare internal ions produced from peptide activation by HCD, we first used sequences identified by merged searches with MS-GFDB of *S. pombe* CID/ETD/HCD triples. The merged searches afforded HCD spectra that were insufficient in themselves for peptide identification. To identify internal ions in the HCD spectra, all possible internal ions from the identified peptide sequence were predicted using an in house program created in [R]. The raw HCD spectrum corresponding to the matched peptide was then searched for the presence of each internal ion. A similar analysis was done on the ddDT spectra to determine the presence of internal ion peaks in the CID spectra from this larger data set. All peptide-spectra matches from ddDT spectra were analyzed for the intensity and presence of *b*-, *y*-, *c*-, *z*-, and internal ions. The *b*- ion count included *b*-H_2O and *b*-NH_3 if the peptide sequence contained serine/threonine, or asparagine/glutamine, and similarly losses were included in the *y*-ion and internal ion counts. Intact precursor ions and neutral losses from the precursors were removed from ETD spectra using the msconvert.exe ETD filter before computing *c*- and *z*- ions and *c*-1 and *z*+1 ions were included in the *c*- and *z*- ion values. The ions were quantified both as %TIC and as % of all MSMS peaks in the spectrum. The fraction of peptide backbone breaks, defined by the presence of a *b*- or *y*-fragment ion corresponding to a break in the peptide backbone, were calculated according to Guthals et al (Guthals and Bandeira 2012).

10. Data analysis

MS-GFDB search output from the activation comparison experiments was filtered to <1% peptide-level FDR. Proteome coverage calculations used only data with <1% peptide-level FDR calculated by PeptideProphet (Keller et al. 2002). Protein identifications were by ProteinProphet with the default parameters (Nesvizhskii et al. 2003). Euler diagrams were generated using eulerAPI (http://www.eulerdiagrams.org/eulerAPE/). Additional analyses were carried out using in-house scripts written in [R] (Team 2011), which have been made available online at github.com/jgmeyerucsd/PepsuM/. Protease specificity heatmaps were generated using only unique peptide sequences from PeptideProphet output. Trans-membrane proteins were predicted from all identified proteins using TMHMM (Krogh et al. 2001). The peptide sequences were analyzed using iceLogos (Colaert et al. 2009).

C. RESULTS

1. Protease activity in SDC, SDS, and GdnCl

Previous studies on WaLP indicated that it possessed remarkable stability (Baker et al. 1992). As this property may provide advantages for digestion of proteome samples under various solution conditions, we performed protease activity assays in various proteomic digestion conditions to assess the versatility of WaLP and MaLP compared to trypsin, LysC and chymotrypsin. In every condition, the activity of WaLP and MaLP was similar or greater compared to trypsin, however chymotrypsin showed higher activity than WaLP in urea and guanidine (Table 3.1A). Strikingly, however, chymotrypsin activity decreased markedly over time under typical proteomic digestion conditions, whereas the activity of WaLP and MaLP remained high (Table

Table 3.1. A. Relative activity in various conditions compared to no denaturant control. **B.** Relative activity of various proteases (as percent compared to time 0) after incubation for 20 hrs.

A.

Condition	Trypsin	Lys C	Chymotrypsin	WaLP	MaLP
No detergent	100	100	100	100	100
0.1% SDC	59	48	89	186	132
1.0% SDC	55	62	130	78	62
0.1% SDS	10	96	19	54	39
1.0% SDS	0	71	0	42	31
1M GdnCl	16	113	83	22	19
4M GdnCl	2	14	16	1	1
1M urea	87	116	82	26	58
4M urea	57	149	57	12	30

B.

Condition	Trypsin	Lys C	Chymotrypsin	WaLP	MaLP
No detergent	64	103	1	28	11
0.1% SDC	87	60	6	96	96
1.0% SDC	85	92	22	99	109

3.1B). These results suggest that for digestion of complex proteomes requiring several hours of digestion, WaLP and MaLP may be superior to chymotrpysin, and may provide a reason for our inability to find reports of complex proteome digestions utilizing chymotrypsin.

2. Coverage of standard protein mixture

A standard protein mixture digested by various proteases was analyzed by FT-CID/ETD/HCD to determine proteome coverage for comparisons to recently published results (Guthals et al. 2013). Digestion of these simple standard protein mixtures gave relatively high protein sequence coverage regardless of the protease (trypsin, LysC, chymotrypsin, elastase, WaLP, or MaLP) (Table 3.2). Compared to all others, chymotrypsin yielded slightly longer peptides on average, whereas elastase yielded slightly shorter peptides. The average length of peptides generated by WaLP and MaLP were similar to trypsin. Similarly high protein sequence coverage was obtained when WaLP and MaLP digest data were combined with trypsin, Lys-C as compared with a recent report using combined data from trypsin, Lys-C, Glu-C, Asp-N, Chymotrypsin, and Arg-C digests of a similar mixture (Table 3.3).

3. Comparison of tryptic and non-tryptic peptide identification using CID, ETD, and HCD

Peptide fragment ion series depend on the peptide amino acid sequence and the activation method used to induce fragmentation (Sleno and Volmer 2004). Tryptic peptides, which bear at least one positive charge at each terminus, produce strong b-

Table 3.2. Percent protein coverage and average peptide lengths obtained from digestion of a standard protein mixture with various proteases.

protein	Trypsin	LysC	Elastase	Chymotrypsin	WaLP	MaLP
GroEL	87.8	92.6	73.7	97.6	80.5	81.0
Leptin	96.5	78.1	52.7	92.5	95.2	80.2
Aprotinin	70.7	91.4	65.5	94.8	96.6	79.3
Peroxidase	58.1	38.6	47.7	64.6	58.5	57.2
BSA	81.8	92.3	79.4	95.7	89.2	83.5
Cyt. C	66.3	67.3	82.7	78.9	82.7	69.3
Hb alpha	81.6	55.3	47.5	66.0	68.8	51.0
Hb beta	86.3	82.9	35.6	91.1	56.2	44.5
Average peptide length	12.5 ± 5.1	15.0 ± 8.1	10.7 ± 3.1	14.8 ± 6.4	11.7 ± 4.8	12.0 ± 4.5

Table 3.3. Percent protein coverage obtained from combining data obtained by digestion of a standard protein mixture with various proteases compared to results from Guthals et al. (23).* Our results and those of Guthals et al., both demonstrate that the mature sequence of Aprotinin given in UniProt is too short, actually it appears to contain an additional N-terminal residue (we found residues 35-93, not 36-93 as reported in UniProt).

Hb beta	Hb alpha	Cyt. C	BSA	Peroxidase	Aprotinin	Leptin	GroEL	protein
86.3	81.6	66.3	81.8	58.1	70.7	96.5	87.8	Trypsin
82.9	55.3	67.3	92.3	38.6	91.4	78.1	92.6	LysC
56.2	68.8	82.7	89.2	58.5	96.6	95.2	80.5	WaLP
44.5	51.0	69.3	83.5	57.2	79.3	80.2	81.0	MaLP
92.5	99.3	85.6	98.6	77.9	102.0*	98.0	98.4	Combined
				67.7	102.0*	94.6	97.8	percent of combined (Guthals et al)
135/146	140/141	89/104	575/583	240/308	59/58	143/146	539/548	combined residues/ residues in protein
1.072	1.218	1.290	1.205	1.341	1.443	1.016	1.121	fold gain of combined/ trypsin

and y-ion series upon activation with CID or HCD. Since peptides from WaLP and MaLP digestion lack such defined charge character, we used the versatile fragmentation ability of the LTQ-Orbitrap Velos (Olsen et al. 2009) equipped with ETD to assess identification efficiency of non-tryptic peptides by CID, ETD, and HCD. First, we analyzed data from total *S. pombe* digests in which peptides identified in MS1 were sequentially activated by CID, ETD, and HCD to compare the results for aLP digests to trypsin digests. This analysis revealed some challenges related to the fact that WaLP and MaLP generate non-tryptic peptides and cleave after several different amino acid residues. Out of the three FT-measured MS/MS activations, FT-HCD was most efficient for identification of tryptic peptides, and the overlap between peptides identified by all three activation methods was high (73%) (Figure3.2). FT-CID and FT-HCD performed similarly for identification of non-tryptic peptides from WaLP digestion and overlap was considerably lower (65%). The greatest overlap of unique identifications was for peptides from Lys-C, with 85% of unique sequences identified by all three activations. Figure 3.3 shows CID, ETD, and HCD spectra for the same peptide from the WaLP digestion. The CID spectrum contains a significant number of peaks due to losses of water and ammonia, and the HCD spectrum contains internal fragment ions, both of which are known to increase spectral complexity (Elias et al. 2004; Huang et al. 2005) resulting in lower PSM scores for peptides that do not have well-defined terminal residues. ETD resulted in abundant charge-reduced precursors along with a low intensity series of c- and z- sequence ions (Fig. 3.2C). ETD contributed a greater percentage of non-overlapping peptide IDs for WaLP and MaLP than for trypsin or LysC, consistent with

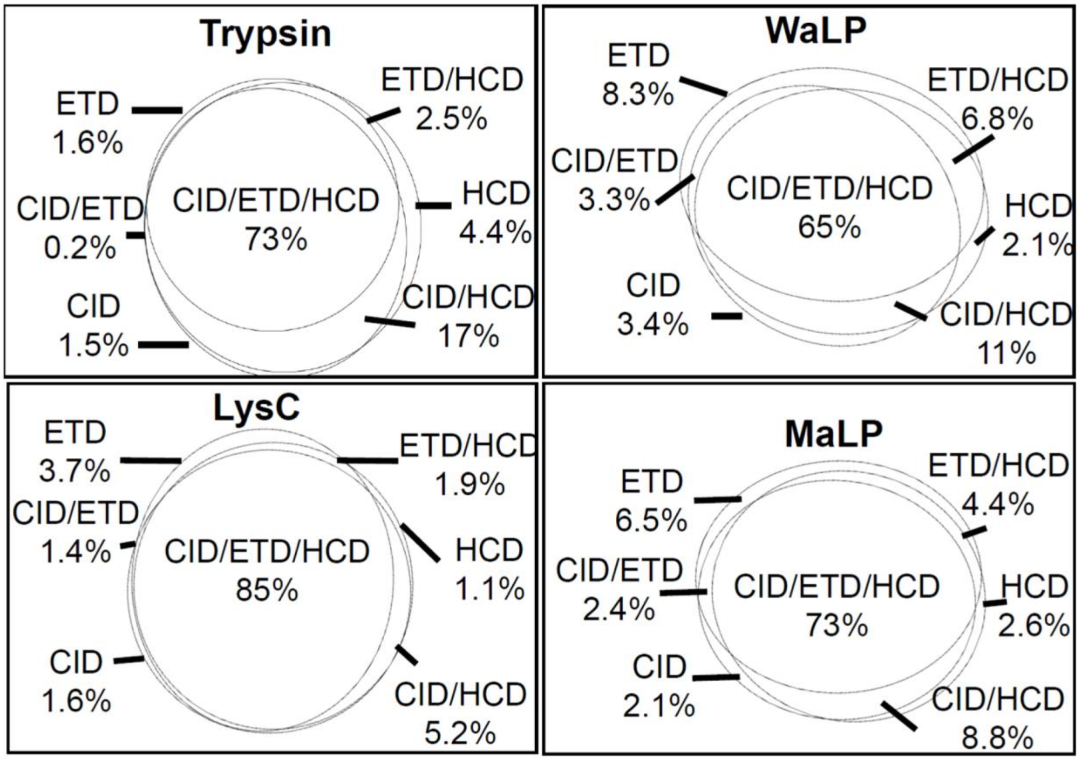


Figure 3.2. Euler diagrams showing the contribution of CID, ETD, and HCD to all unique peptides identified from sequential activation analyses of unfractionated *S. pombe* proteome analyzed over a three hour reverse phase separation (resulting in a total of 1836 peptides identified from the trypsin digest, 1307 from the LysC digest, 1744 from the WaLP digest and 1327 from the MaLP digest). The greatest overlap in identifications was observed for peptides from LysC digestion, and the least overlap was observed for peptides from WaLP digestion. ETD contributes a greater percentage of unique sequences for peptides from WaLP and MaLP digestion relative to peptides from trypsin or LysC digestion.

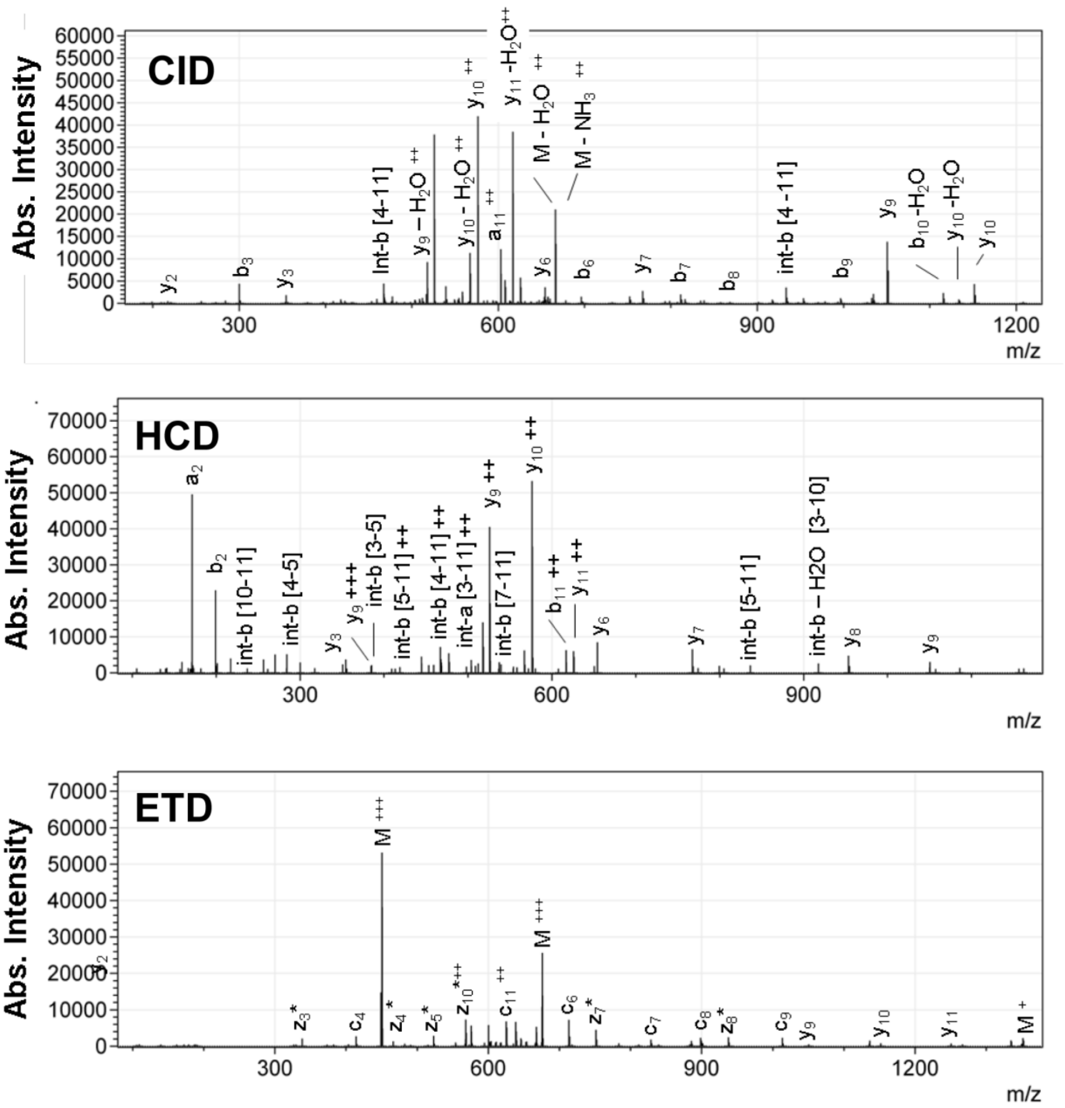


Figure 3.3. CID, HCD, and ETD spectra of a peptide from WaLP digestion: VVTPWLDGKHVV CID and HCD (M+2H=675.3824), ETD (M+3H=450.5907).

our previous study comparing CID and ETD for non-tryptic peptides from elastase and pepsin (Meyer and Komives 2012).

We re-searched these spectra using a merged search protocol in MS-GFDB. This approach resulted in more peptide identifications within 1% FDR. Merged searching resulted in only marginal improvements in the number of identifications of peptides from the trypsin and LysC digests (5% increases) but larger improvements (Table 3.4) were obtained for the samples from MaLP and WaLP digests. This is not surprising since these searches were run without enzyme specificity and thus have most to gain when capitalizing on the CID/ETD/HCD complementarity when searching a larger search space. The merged search data also allowed us to analyze the HCD spectra that were not sufficient for peptide identification in the absence of additional information from ETD and/or CID (56% of the triples). We first analyzed all of the HCD spectra for identified peptides to determine the percentage of the total ion current (%TIC) contributed by internal ions (7.2 % for trypsin, 7.4% for LysC vs. 9.2% for WaLP and 9.1% for MaLP). A statistically significant greater percent of the TIC was contributed by internal ions from WaLP and MaLP digests as compared to trypsin digestion (student's t-test, p-values $<10^{-10}$). We next analyzed the fraction of MSMS peaks attributable to internal ions. In this analysis, only peptides from the WaLP digest yielded a statistically significant increase in the fraction of MSMS peaks attributable to internal ions (10.3%), whereas the relative number of internal ion peaks from the MaLP digestion (8.4%) was similar to that of trypsin (8.7%), and peptides from LysC produced slightly fewer internal ion peaks (7.9%). Using all peptides identified from the four separate digestions, we examined the cross-correlation of

Table 3.4. Comparison of the number of unique peptide sequences identified from CID/HCD/ETD data from 3 hr nLC MS analyses of *S. pombe* proteome digests.

	MSGFDB separate	MSGFDB merged	Protein Prospector	gain from merged search
Trypsin	1771	1839	1836	1.04
LysC	1012	1064	1307	1.05
WaLP	1169	1520	1744	1.30
MaLP	737	971	1327	1.32

internal ion abundance with the presence of each of the 20 amino acids. The abundance and presence of internal ions was positively correlated with the presence of residues: A, D, G, I, L, P, and V. Interestingly, only arginine was found to negatively correlate to internal ions.

We next used data from the CID/ETD activation comparison to determine branch points for a data-dependent decision tree (ddDT), which targets precursors for CID or ETD based on precursor charge state and m/z (figure 3.4). A ddDT targeting +2 charge-state precursors with CID, and ≥+3 charge-state precursors with ETD was implemented similarly to previous reports (Swaney et al. 2008). The total run time of the ddDT method was only 1.5 hours, half that of the other activation comparison runs. Use of this ddDT afforded more unique peptide identifications from WaLP digestion than even the best 3 hour activation experiment (i.e. 2544 from 1.5 hr. ddDT versus 2358 from merged search of IT-CID/ETD). Use of the ddDT for tryptic peptides resulted in nearly as many peptides as IT-CID in half the acquisition time (i.e. 4195 from 1.5 hr. IT-ddDT, 4576 from 3hr. IT-CID).

4. Characterization of the MSMS data from WaLP and MaLP digests

S. pombe proteome samples digested separately by trypsin, LysC, WaLP or MaLP were separated off-line using HPRP (Wang et al. 2011), and each fraction was analyzed using a 90 minute nLC run with the ddDT method. Over 200,000 MSMS spectra were collected for each sample and searched with MSGF+ followed by training and re-searching; at less than 1% peptide-level FDR, similar numbers of peptides were identified from the trypsin, LysC, WaLP and MaLP digests; 17,810 and 26,747 peptides were identified from the WaLP and MaLP digests, respectively (Table

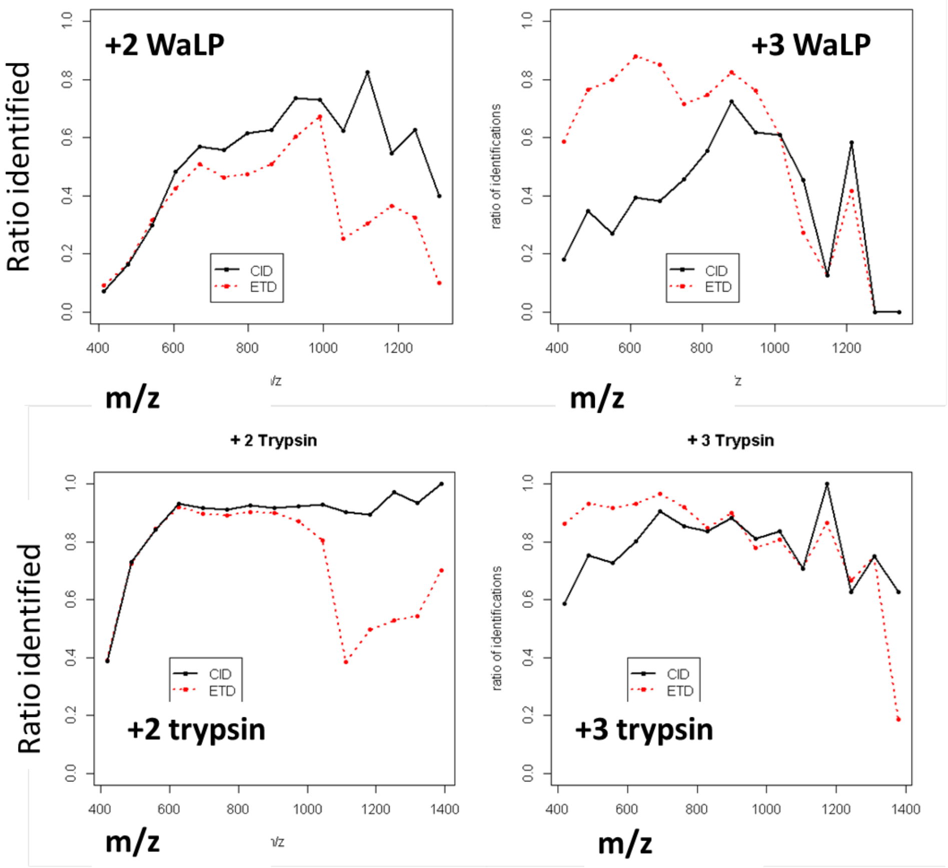


Figure 3.4. Plots of identification efficiency as a function of precursor mass/charge separated by precursor charge state. Identification efficiency of peptides from WaLP digestion are plotted on the top, and identification efficiency of peptides from trypsin digestion are plotted on the bottom.

Table 3.5. Results for proteome coverage based on unique peptides for each individual digest and combined data from different digests.

Data set	Spectra	Peptides	Protein groups	Proteome Coverage
Trypsin (T)	246,996	20,480	2,837	9.2%
LysC (L)	226,403	21,565	2,781	8.4%
WaLP (W)	267,608	17,810	1,955	5.8%
MaLP (M)	251,103	26,747	2,330	7.6%
T1+T2*		23,069	2,947(+6%)	10.2% (+10%)
T+L		37,808	3,293(+16%)	13.6% (+48%)
T+W		38,282	3,007 (+6%)	12.5% (+36%)
T+M		47,111	3,207 (+13%)	13.9% (+51%)
T+L+W+M		79,508	3,555 (+24%)	18.5% (+101%)

3.5). Even though a very similar number of spectra were collected for each digest, the number of peptides identified from the WaLP digest was somewhat lower. Several factors might have contributed to this, one being the non-tryptic C-termini generated by this enzyme.

To better understand the consequences of non-tryptic C-termini on peptide identification, we analyzed the number of various fragment ions observed in the MSMS spectra from trypsin, LysC, WaLP or MaLP digests (Table 3.6). The percentage of the TIC attributable to y-ions is higher for trypsin and LysC than for WaLP and MaLP and conversely the percentage of the TIC attributable to b-ions is higher for WaLP and MaLP (Figure3.5). The y-ion directing capabilities of C-terminal positive charges has been discussed previously (Tabb et al. 2004), but the impact on large proteome analyses can be appreciated from the results presented here. Indeed, many search engines give higher scores for y ions than for other ions, which may be part of the reason more peptides were identified from the tryptic digestion than from the WaLP digestion. Another possible reason for lower numbers of peptide IDs from the WaLP digest could be production of internal ions upon MS/MS. Training the MSGF+ scoring function for peptides from WaLP/MaLP allowed reasonable identification of non-tryptic peptides despite the lower percentage of the TIC attributable to y-ions. In fact, the MaLP digested sample resulted in the greatest number of unique peptides identified (Table 3.5).

5. Substrate specificity of WaLP and MaLP

Previous studies of WaLP specificity were based on chromogenic activity assays revealed high activity towards P1 residues: A, V, and M

Table 3.6. Analysis of ion types observed from peptides resulting from digestion with various proteases.

Protease	subset	% TIC	% MS/MS peaks	Notes
Trypsin	CID, b	14.8 ± 8.6%	8.30 ± 3.4%	CID PSMs = 65,891
	CID, y	20.9 ± 10.7%	9.0 ± 5.7%	
	CID breaks	81.8 ± 15.3%		
	ETD, c	21.1 ± 8.4%	12.8 ± 4.6%	ETD PSMs = 13,222
	ETD, z	23.0 ± 8.6%	14.1 ± 5.6%	
	ETD breaks	47.5 ± 26.0%		
LysC	CID, b	17.2 ± 10.3 %	7.0 ± 3.0%	CID PSMs = 48,742
	CID, y	18.2 ± 10.7%	6.9 ± 2.8 %	
	CID breaks	83.7 ± 12.7%		
	ETD, c	20.2 ± 8.4%	8.1 ± 3.9%	ETD PSMs = 19,738
	ETD, z	22.1 ± 8.5%	9.0 ± 4.2%	
	ETD breaks	64.75 ± 25.0%		
WaLP	CID, b	22.4 ± 11.1 %	7.5 ± 3.1 %	CID PSMs = 33,475
	CID, y	14.5 ± 9.0%	6.7 ± 2.6%	
	CID breaks	80.4 ± 12.1%		
	ETD, c	23.3 ± 8.8 %	10.3 ± 4.6 %	ETD PSMs = 11,117
	ETD, z	19.1 ± 7.8 %	9.5 ± 4.0%	
	ETD breaks	53.5 ± 23.6%		
MaLP	CID, b	20.1 ± 10.5%	6.08 ± 2.44%	CID PSMs = 45,138
	CID, y	14.4 ± 9.29%	5.82 ± 2.13%	
	CID breaks	84.3 ± 11.9%		
	ETD, c	22.2 ± 8.61%	7.05 ± 3.68%	ETD PSMs = 13,509
	ETD, z	18.5 ± 7.59%	6.55 ± 3.09%	
	ETD breaks	69.9 ± 20.6%		

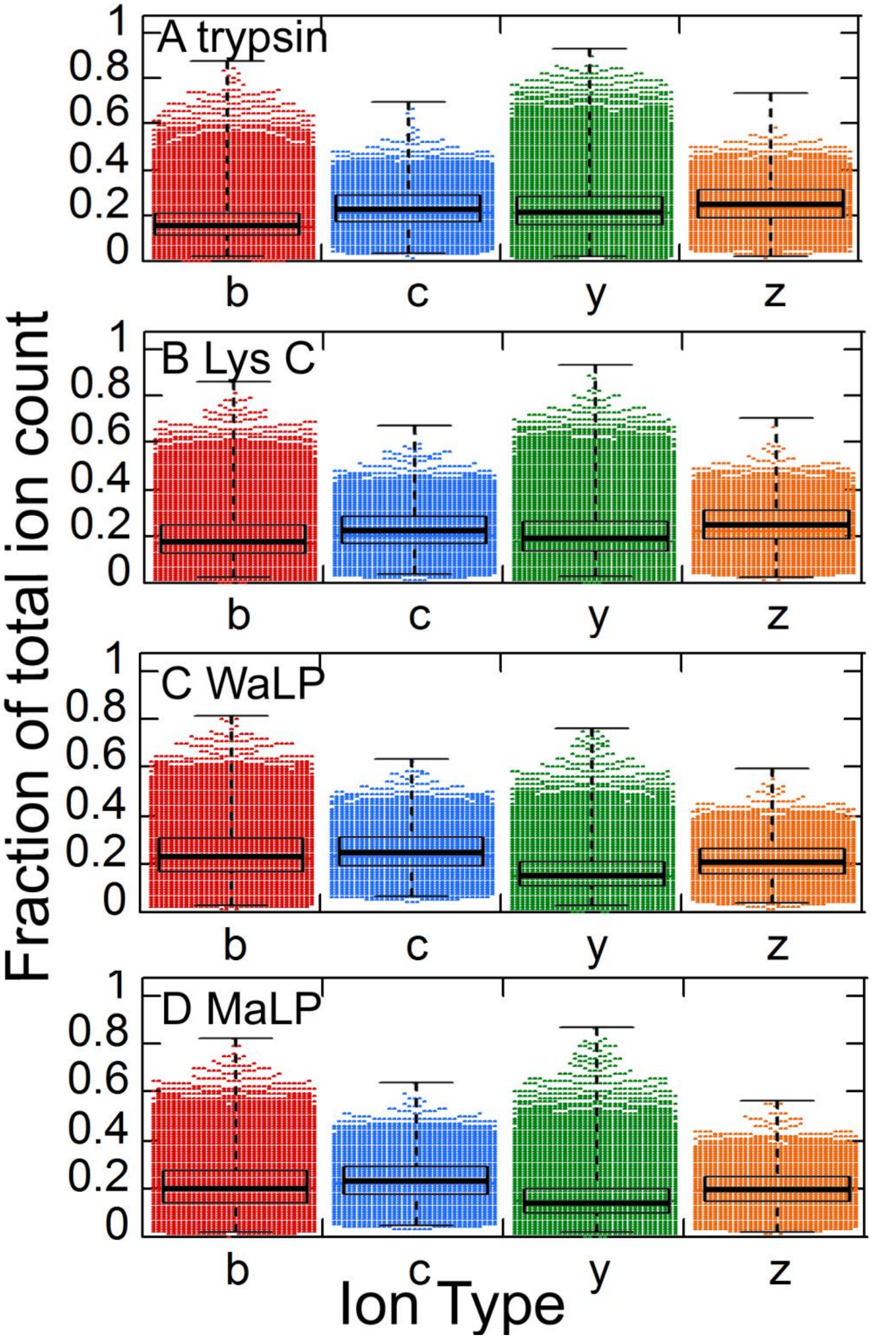
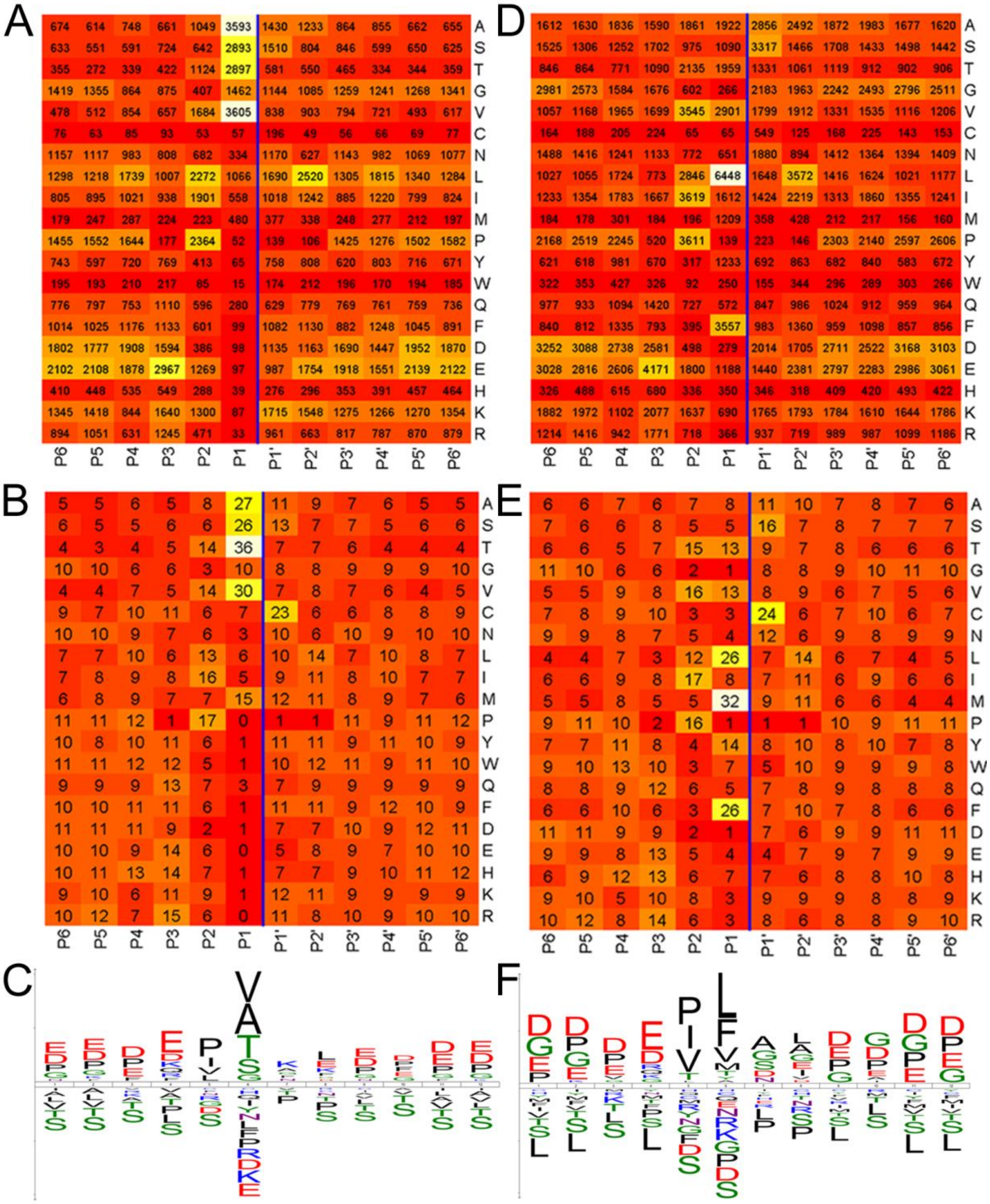


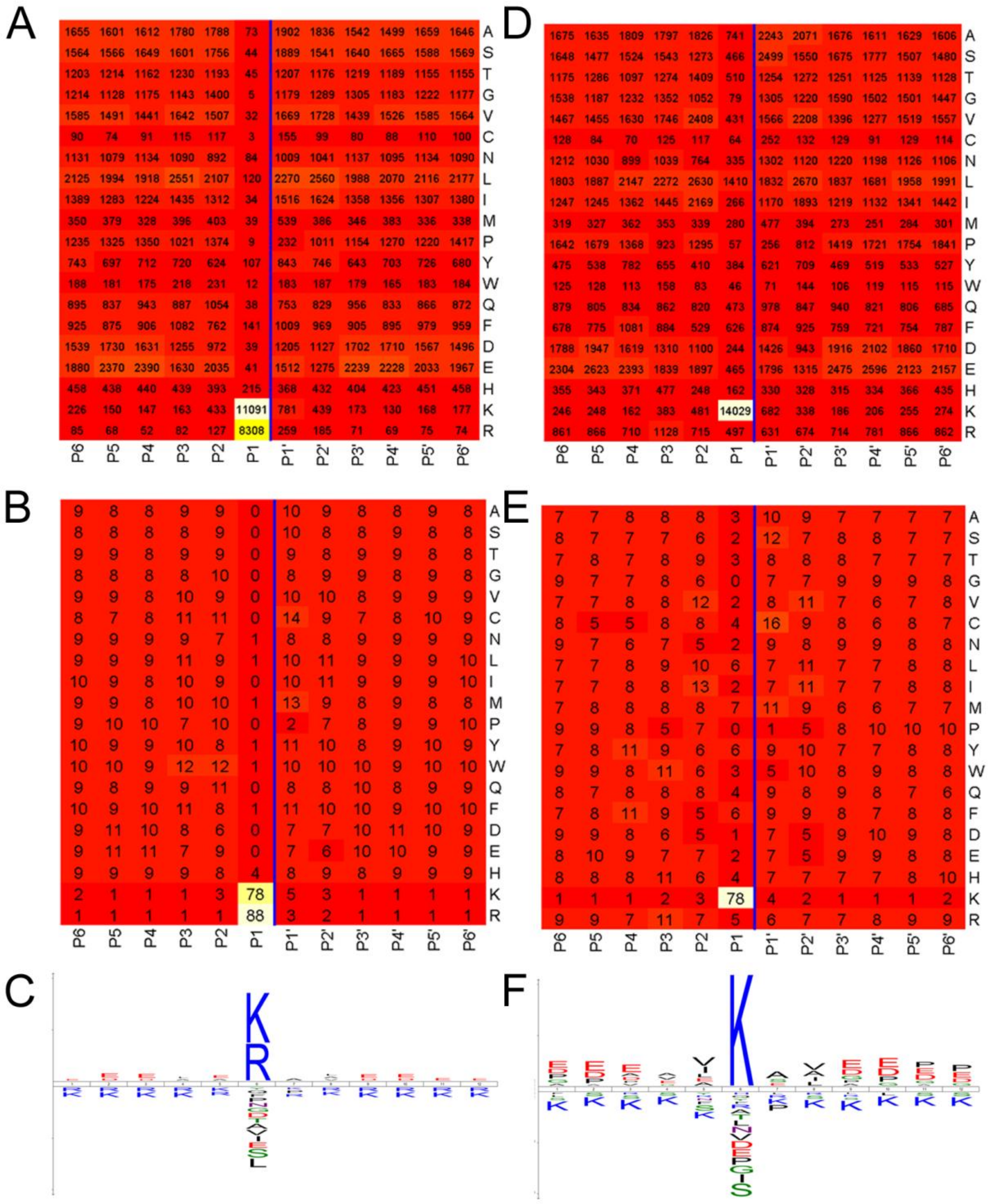
Figure 3.5. The fraction of observed ion types in the MSMS spectra from which peptides could be identified from the datasets detailed in Table 3.4 (A, trypsin; B, LysC; C, WaLP; D, MaLP; over 17,000 peptides from the WaLP digest and over 26,000 peptides from the MaLP digest). The black bars represent the inner quartile ranges and show that for WaLP and MaLP digests, a significantly smaller proportion of the TIC was accounted for by C-terminal ions (y-ions and z-ions) as compared to trypsin and LysC digests.

(Bone et al. 1989). MaLP, which has an active site Met replaced by Ala, was reported to have broadened activity for M, L, and F, but similar activity still against A and V (Bone et al. 1989). To more fully characterize the substrate specificity, all unique peptide sequences from PeptideProphet were combined to determine the specificity of WaLP and MaLP. The observed specificity of WaLP and MaLP were visualized by plotting heat maps of cleavage position and observed amino acid frequency (Figure3.6). For comparison, the same figure is given for peptides from trypsin and LysC (Figure 3.7). WaLP cleaves most frequently after T (36%), but also with significant frequency after V (30%), A (27%), S (26%), and M (16%). As reported previously, MaLP has specificity for slightly larger aliphatic amino acids, cleaving most frequently after M (32%), L (26%), F (26%), Y (14%), T (13%) and V (13%). These results show that WaLP and MaLP are somewhat more specific than elastase, which cleaved after A (43.5%), V (36.5%), I (34.7%), T (30.3%), S (21.4%), L (19.5%), and M (15.7%) (Wang et al. 2008). Interestingly, MaLP appears to be able to differentiate between L and I (26% of leucines were found at the P1 position versus only 8% of isoleucines), which cannot be resolved by mass alone. To follow up this potentially very interesting finding, we measured the ability of MaLP to cleave succinyl-A-A-P-L-pNa vs. succinyl-A-A-P-I-pNa. Whereas activity towards the succinyl-A-A-P-L-pNa was high (specific activity of 3.4x10-3 U/mg compared to 2.8x10-2 U/mg for the substrate of choice, succinyl-A-A-P-F-pNa), the activity of MaLP towards succinyl-A-A-P-I-pNa was not observable under identical assay conditions.

6. Length of peptides from WaLP and MaLP digestions

Figure 3.6. Heat maps (with white representing the highest and red representing the lowest) summarizing the statistical analysis of the occurrence of each amino acid at each position in the identified peptides detailed in Table 3.5 (over 17,000 peptides from the WaLP digest and over 26,000 peptides from the MaLP digest). (A) WaLP peptides, raw counts of frequency of each amino acid at each position (B) WaLP peptides, counts normalized for the occurrence of each amino acid at each position (C) IceLogo depicting the enrichment and depletion of specific amino acids (relative to the whole proteome) at each position in the WaLP peptides with residues colored according to property: acidic-red, basic-blue, hydrophobic-black and small/neutral-green. WaLP yields peptides with the following P1 (C-terminal) residues A(20%), V (20%), S(16%), T(16%), G(8%), L(6%). (D) MaLP peptides, raw counts of frequency of each amino acid at each position (E) MaLP peptides, counts normalized for the occurrence of each amino acid at each position. (F) IceLogo depicting the enrichment and depletion of specific amino acids (relative to the whole proteome) at each position in the MaLP peptides. MaLP yields peptides with the following P1 (C-terminal) residues L(24%), F(13%), V(11%), A(7%), T(7%), I(6%). The cleavage site is marked by the vertical line in each plot.


A
674 614 748 661 1049 3593 1430 1233 864 855 662 655 A
633 551 591 724 642 2893 1510 804 846 599 650 625 S
355 272 339 422 1124 2897 581 550 465 334 344 359 T
1419 1355 864 875 407 1462 1144 1085 1259 1241 1268 1341 G
478 512 854 657 1684 3605 838 903 794 721 493 617 V
76 63 85 93 53 57 196 49 56 66 69 77 C
1157 1117 983 808 682 334 1170 627 1143 982 1069 1077 N
1298 1218 1739 1007 2272 1066 1690 2520 1305 1815 1340 1284 L
805 895 1021 938 1901 558 1018 1242 885 1220 799 824 I
179 247 287 224 223 480 377 338 248 277 212 197 M
1455 1552 1644 177 2364 52 139 106 1425 1276 1502 1582 P
743 597 720 769 413 65 758 808 620 803 716 671 Y
195 193 210 217 85 15 174 212 196 170 194 185 W
776 797 753 1110 596 280 629 779 769 761 759 736 Q
1014 1025 1176 1133 601 99 1082 1130 882 1248 1045 891 F
1802 1777 1908 1594 386 98 1135 1163 1690 1447 1952 1870 D
2102 2108 1878 2967 1269 97 987 1754 1918 1551 2139 2122 E
410 448 535 549 288 39 276 296 353 391 457 464 H
1345 1418 844 1640 1300 87 1715 1548 1275 1266 1270 1354 K
894 1051 631 1245 471 33 961 663 817 787 870 879 R
P6 P5 P4 P3 P2 P1 P1' P2' P3' P4' P5' P6'
D
1612 1630 1836 1590 1861 1922 2856 2492 1872 1983 1677 1620 A
1525 1306 1252 1702 975 1090 3317 1466 1708 1433 1498 1442 S
846 864 771 1090 2135 1959 1331 1061 1119 912 902 906 T
2981 2573 1584 1676 602 266 2183 1963 2242 2493 2796 2511 G
1057 1168 1965 1699 3545 2901 1799 1912 1331 1535 1116 1206 V
164 188 205 224 65 65 549 125 168 225 143 153 C
1488 1416 1241 1133 772 651 1880 894 1412 1364 1394 1409 N
1027 1055 1724 773 2846 6448 1648 3572 1416 1624 1021 1177 L
1233 1354 1783 1667 3619 1612 1424 2219 1313 1860 1355 1241 I
184 178 301 184 196 1209 358 428 212 217 156 160 M
2168 2519 2245 520 3611 139 223 146 2303 2140 2597 2606 P
621 618 981 670 317 1233 692 863 682 840 583 672 Y
322 353 427 326 92 250 155 344 296 289 303 266 W
977 933 1094 1420 727 572 847 986 1024 912 959 964 Q
840 812 1335 793 395 3557 983 1360 959 1098 857 856 F
3252 3088 2738 2581 498 279 2014 1705 2711 2522 3168 3103 D
3028 2816 2606 4171 1800 1188 1440 2381 2797 2283 2986 3061 E
326 488 615 680 336 350 346 318 409 420 493 422 H
1882 1972 1102 2077 1637 690 1765 1793 1784 1610 1644 1786 K
1214 1416 942 1771 718 366 937 719 989 987 1099 1186 R
P6 P5 P4 P3 P2 P1 P1' P2' P3' P4' P5' P6'
B
5 5 6 5 8 27 11 9 7 6 5 5 A
6 5 5 6 6 26 13 7 7 5 6 6 S
4 3 4 5 14 36 7 7 6 4 4 4 T
10 10 6 6 3 10 8 8 9 9 9 10 G
4 4 7 5 14 30 7 8 7 6 4 5 V
9 7 10 11 6 7 23 6 6 8 8 9 C
10 10 9 7 6 3 10 6 10 9 10 10 N
7 7 10 6 13 6 10 14 7 10 8 7 L
7 8 9 8 16 5 9 11 8 10 7 7 I
6 8 9 7 7 15 12 11 8 9 7 6 M
11 11 12 1 17 0 1 1 11 9 11 12 P
10 8 10 11 6 1 11 11 9 11 10 9 Y
11 11 12 12 5 1 10 12 11 9 11 10 W
9 9 9 13 7 3 7 9 9 9 9 9 Q
10 10 11 11 6 1 11 11 9 12 10 9 F
11 11 11 9 2 1 7 7 10 9 12 11 D
10 10 9 14 6 0 5 8 9 7 10 10 E
10 11 13 14 7 1 7 7 9 10 11 12 H
9 10 6 11 9 1 12 11 9 9 9 9 K
10 12 7 15 6 0 11 8 10 9 10 10 R
P6 P5 P4 P3 P2 P1 P1' P2' P3' P4' P5' P6'
E
6 6 7 6 7 8 11 10 7 8 7 6 A
7 6 6 8 5 5 16 7 8 7 7 7 S
6 6 5 7 15 13 9 7 8 6 6 6 T
11 10 6 6 2 1 8 8 9 10 11 10 G
5 5 9 8 16 13 8 9 6 7 5 6 V
7 8 9 10 3 3 24 6 7 10 6 7 C
9 9 8 7 5 4 12 6 9 8 9 9 N
4 4 7 3 12 26 7 14 6 7 4 5 L
6 6 9 8 17 8 7 11 6 9 6 6 I
5 5 8 5 5 32 9 11 6 6 4 4 M
9 11 10 2 16 1 1 1 10 9 11 11 P
7 7 11 8 4 14 8 10 8 10 7 8 Y
9 10 13 10 3 7 5 10 9 9 9 8 W
8 8 9 12 6 5 7 8 9 8 8 8 Q
6 6 10 6 3 26 7 10 7 8 6 6 F
11 11 9 9 2 1 7 6 9 9 11 11 D
9 9 8 13 5 4 4 7 9 7 9 9 E
6 9 12 13 6 7 7 6 8 8 10 8 H
9 10 5 10 8 3 9 9 9 8 8 9 K
10 12 8 14 6 3 8 6 8 8 9 10 R
P6 P5 P4 P3 P2 P1 P1' P2' P3' P4' P5' P6'
C
F



A

	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'
A	1655	1601	1612	1780	1788	73	1902	1836	1542	1499	1659	1646
S	1564	1566	1649	1601	1756	44	1889	1541	1640	1665	1588	1569
T	1203	1214	1162	1230	1193	45	1207	1176	1219	1189	1155	1155
G	1214	1128	1175	1143	1400	5	1179	1289	1305	1183	1222	1177
V	1585	1491	1441	1642	1507	32	1669	1728	1439	1526	1585	1564
C	90	74	91	115	117	3	155	99	80	88	110	100
N	1131	1079	1134	1090	892	84	1009	1041	1137	1095	1134	1090
L	2125	1994	1918	2551	2107	120	2270	2560	1988	2070	2116	2177
I	1389	1283	1224	1435	1312	34	1516	1624	1358	1356	1307	1380
M	350	379	328	396	403	39	539	386	346	383	336	338
P	1235	1325	1350	1021	1374	9	232	1011	1154	1270	1220	1417
Y	743	697	712	720	624	107	843	746	643	703	726	680
W	188	181	175	218	231	12	183	187	179	165	183	184
Q	895	837	943	887	1054	38	753	829	956	833	866	872
F	925	875	906	1082	762	141	1009	969	905	895	979	959
D	1539	1730	1631	1255	972	39	1205	1127	1702	1710	1567	1496
E	1880	2370	2390	1630	2035	41	1512	1275	2239	2228	2033	1967
H	458	438	440	439	393	215	368	432	404	423	451	458
K	226	150	147	163	433	11091	781	439	173	130	168	177
R	85	68	52	82	127	8308	259	185	71	69	75	74

D

	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'
A	1675	1635	1809	1797	1826	741	2243	2071	1676	1611	1629	1606
S	1648	1477	1524	1543	1273	466	2499	1550	1675	1777	1507	1480
T	1175	1286	1097	1274	1409	510	1254	1272	1251	1125	1139	1128
G	1538	1187	1232	1352	1052	79	1305	1220	1590	1502	1501	1447
V	1467	1455	1630	1746	2408	431	1566	2208	1396	1277	1519	1557
C	128	84	70	125	117	64	252	132	129	91	129	114
N	1212	1030	899	1039	764	335	1302	1120	1220	1198	1126	1106
L	1803	1887	2147	2272	2630	1410	1832	2670	1837	1681	1958	1991
I	1247	1245	1362	1445	2169	266	1170	1893	1219	1132	1341	1442
M	319	327	362	353	339	280	477	394	273	251	284	301
P	1642	1679	1368	923	1295	57	256	812	1419	1721	1754	1841
Y	475	538	782	655	410	384	621	709	469	519	533	527
W	125	128	113	158	83	46	71	144	106	119	115	115
Q	879	805	834	862	820	473	978	847	940	821	806	685
F	678	775	1081	884	529	626	874	925	759	721	754	787
D	1788	1947	1619	1310	1100	244	1426	943	1916	2102	1860	1710
E	2304	2623	2393	1839	1897	465	1796	1315	2475	2596	2123	2157
H	355	343	371	477	248	162	330	328	315	334	366	435
K	246	248	162	383	481	14029	682	338	186	206	255	274
R	861	866	710	1128	715	497	631	674	714	781	866	862

B

	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'
A	9	8	8	9	9	0	10	9	8	8	9	8
S	8	8	8	8	9	0	10	8	8	9	8	8
T	9	9	8	9	9	0	9	8	9	9	8	8
G	8	8	8	8	10	0	8	9	9	8	9	8
V	9	9	8	10	9	0	10	10	8	9	9	9
C	8	7	8	11	11	0	14	9	7	8	10	9
N	9	9	9	9	7	1	8	8	9	9	9	9
L	9	9	9	11	9	1	10	11	9	9	9	10
I	10	9	8	10	9	0	10	11	9	9	9	10
M	9	9	8	10	10	1	13	9	8	9	8	8
P	9	10	10	7	10	0	2	7	8	9	9	10
Y	10	9	9	10	8	1	11	10	8	9	10	9
W	10	10	9	12	12	1	10	10	10	9	10	10
Q	9	8	9	9	11	0	8	8	10	8	9	9
F	10	9	10	11	8	1	11	10	10	9	10	10
D	9	11	10	8	6	0	7	7	10	11	10	9
E	9	11	11	7	9	0	7	6	10	10	9	9
H	9	9	9	9	8	4	8	9	8	9	9	9
K	2	1	1	1	3	78	5	3	1	1	1	1
R	1	1	1	1	1	88	3	2	1	1	1	1

E

	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'
A	7	7	8	8	8	3	10	9	7	7	7	7
S	8	7	7	7	6	2	12	7	8	8	7	7
T	7	8	7	8	9	3	8	8	8	7	7	7
G	9	7	7	8	6	0	7	7	9	9	9	8
V	7	7	8	8	12	2	8	11	7	6	7	8
C	8	5	5	8	8	4	16	9	8	6	8	7
N	9	7	6	7	5	2	9	8	9	9	8	8
L	7	7	8	9	10	6	7	11	7	7	8	8
I	7	7	8	8	13	2	7	11	7	7	8	8
M	7	8	8	8	8	7	11	9	6	6	7	7
P	9	9	8	5	7	0	1	5	8	10	10	10
Y	7	8	11	9	6	6	9	10	7	7	8	8
W	9	9	8	11	6	3	5	10	8	9	8	8
Q	8	7	8	8	8	4	9	8	9	8	7	6
F	7	8	11	9	5	6	9	9	8	7	8	8
D	9	9	8	6	5	1	7	5	9	10	9	8
E	8	10	9	7	7	2	7	5	9	9	8	8
H	8	8	8	11	6	4	7	7	7	7	8	10
K	1	1	1	2	3	78	4	2	1	1	1	2
R	9	9	7	11	7	5	6	7	7	8	9	9

Figure 3.7. Heat maps from trypsin (A-C) and Lys C (D-F) data collected at the same time as those shown in Figure 3.5 for MaLP and WaLP digests. Figures A and D show the raw counts of frequency of each amino acid at each position, Figures B and E show the counts normalized for the occurrence of each amino acid at each position, Figures C And F show the IceLogos depicting the enrichment and depletion of specific amino acids (relative to the whole proteome).

Because of their apparently promiscuous activity, complete WaLP or MaLP digestion could result in many single amino acids and short peptides. Remarkably,WaLP digestion produces peptides with nearly the same average length as trypsin, 11.8 ± 3.1 amino acids versus 12.2 ± 4.3 from WaLP and trypsin, respectively (Figure3.8). MaLP digestion produced slightly longer peptides (12.6 ± 3.7 amino acids). In addition, if non-specific digestion was resulting in more single amino acids produced, one would expect that the yield of amino acids still in peptides that adhered to C18 during solid phase extraction would be less, but this was not the case. Amino acid analysis of peptides from each digest revealed similar total peptide yield from digestion by trypsin, WaLP, and MaLP (Figure 3.9). Interestingly, these results suggest that amino acids corresponding to the P1 specificity are depleted. So, for example, the peptides isolated from trypsin digests contain less R and K than the whole proteome and the peptides isolated from WaLP digests contain less A, S, T and V than the whole proteome. This makes sense if pairs of these residues were cleaved into individual amino acids and not retained as peptides in the experiment. Indeed, in silico trypsin cleavage yields digestion products of which nearly one quarter would be only a single amino acid (presumably K or R).

7. Quantitation of peptide overlap

Another possible limitation due to proteome digestion by semi-specific proteases is production of largely redundant sequences with different terminal truncations ("shredding"). We quantified the redundancy in amino acid coverage according to the following relationship: $redundancy_{total} = \frac{\sum \# A.A._{observed}}{\sum \# A.A._{unique}}$

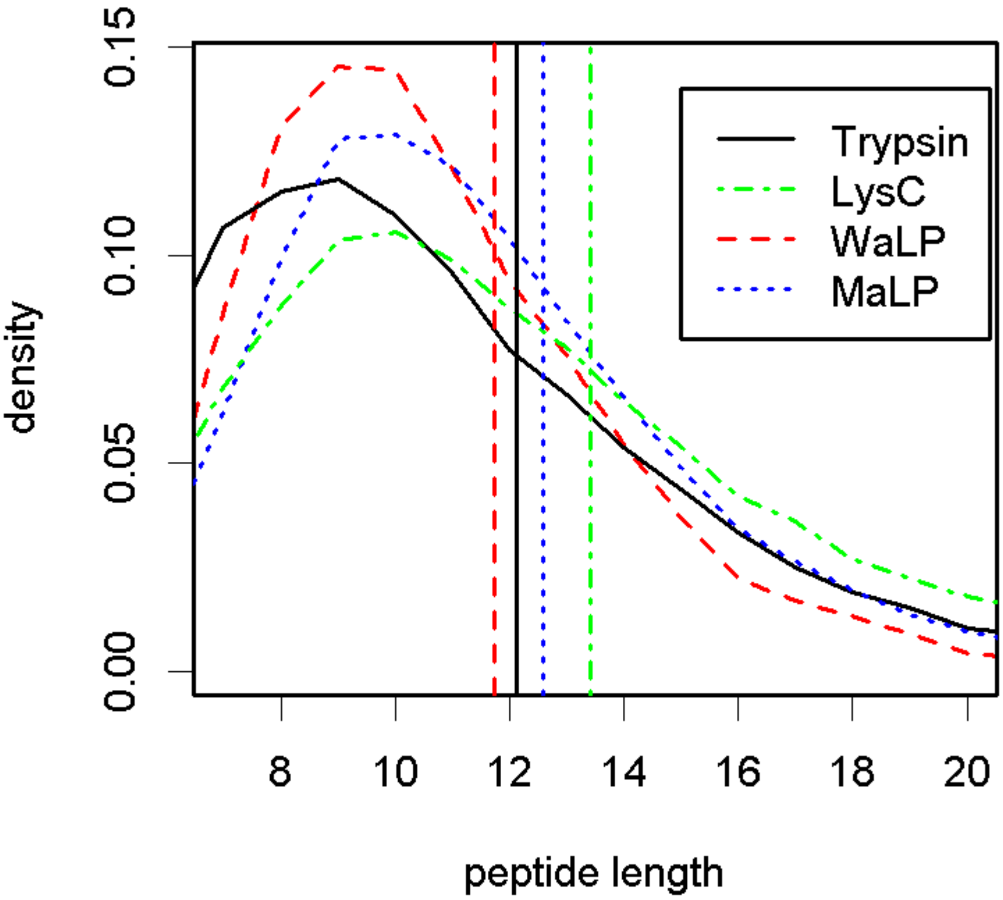


Figure 3.8. Analysis of the lengths of all the unique peptides observed in each of the protease digests from the data in Table 3.5. Trypsin digestion generated a broader distribution with a higher frequency of shorter peptides. The size distributions of the peptides from the WaLP and MaLP digests were narrower than those for either trypsin or LysC. Colored vertical lines mark the average observed peptide length. WaLP digestion produced the shortest mean peptide length of 11.8 ± 3.1 amino acids. Trypsin digestion produced peptides with an average length of 12.2 ± 4.2 amino acids. MaLP and LysC produced slightly longer average peptides with length 12.6 ± 3.7 and 13.5 ± 4.7 amino acids, respectively. The average lengths of the observed peptides were all remarkably similar.

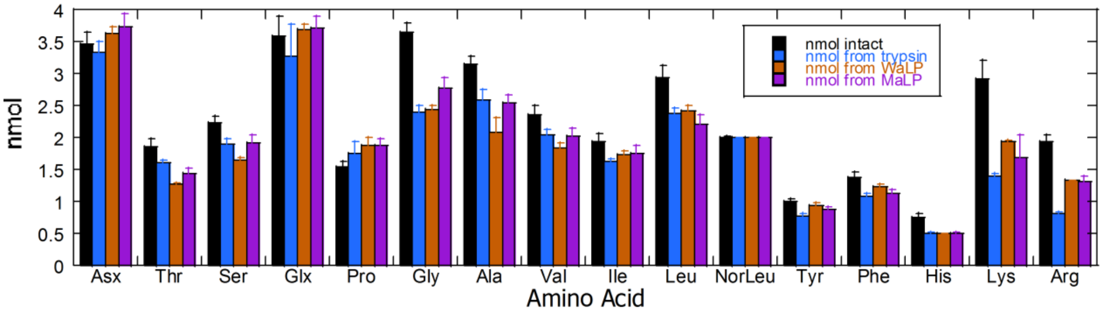


Figure 3.9. Quantitative amino acid analysis of undigested (black), trypsin digested (blue), WaLP digested (orange) and MaLP digested (purple) of *S. pombe* proteome after purification on C18 to remove single amino acids and undigested proteins. The results show that each protease yields similar total amounts of peptides with similar amino acid compositions, however some interesting differences in amino acid content were also discovered (discussed further in the text).

The numerator includes redundancy from chemical modification (e.g. oxidized methionine), overlapping peptides, and identification of multiple charge states. This relationship can be applied to any proteomics experiment to assess the efficiency of converting peptide identifications to covering proteome sequences. Peptides from trypsin digestion were the least redundant, and peptides from MaLP digestion were the most redundant. Redundancy values for trypsin, LysC, WaLP and MaLP were 1.3, 1.6, 1.7, and 2.0, respectively. The redundancy for combined data was 2.7. Such high redundancy is expected to be useful for high ion coverage that would facilitate site localization of PTMs (Chalkley and Clauser 2012).

8. Biological gains from WaLP and MaLP digestions

The central aim of this study was to improve proteome coverage. Compared to data from only trypsin, the combined proteases increased protein identifications by 24%, and proteome coverage by 101% (Table 3.5). Such gains were significantly greater than those afforded from re-injection of tryptic peptides, which only increased proteome coverage by 10%.

One possible gain from the increased proteome coverage would be in PTM identification. Although the samples were not enriched for phosphorylation we re-searched the fully separated *S. pombe* proteome data allowing for variable phosphorylation of S and T to look for these PTMs. Indeed, the complementary amino acids covered from WaLP digestion allowed observance of 95 serine and threonine phosphorylations, 63 of which had not been previously reported in Uniprot. Similarly, 77 S/T phosphorylations were identified from the MaLP digest, 57 of which had not been previously reported (the assignments were made at the peptide level, not the site level). A particularly illustrative example of the improved coverage of phosphorylation sites was observed for the protein MPD2. The WaLP digest contained three phosphorylated peptides from MPD2, one corresponding to the previously reported phosphorylation of S175, and two novel sites at S223 and S750 (Figure 3.10). It is clear from the sequence of the protein that S175 is located between two basic residues and would result in a 17 amino acid long tryptic peptide. The tryptic peptide covering S223 would be 75 amino acids in length, and the tryptic peptide covering S750 would be only five amino acids in length.

We also wondered whether proteases that cleave aliphatic residues might increase the coverage of membrane protein sequences. Out of all 3,555 protein groups identified, 244 (6.9%) were predicted to have 3 or more transmembrane helices. Sequences from these proteins were preferentially enriched in the gained coverage with increases up to 350% for very hydrophobic sequences (Figure3.11). Peptides from MaLP digestion were the greatest individual contributor to these gains as can be seen by the observation that the percent proteome coverage does not decrease with transmembrane helix content nearly as dramatically as for the other proteases (Fig. 3.10A). Since in-gel digestion is sometimes used to digest membrane proteins, we tested the suitability of WaLP and MaLP for in-gel digestion on glucose transporter 5(UniProt Acc# P22732). Trypsin digestion yielded 36% coverage, WaLP yielded 84% coverage, and MaLP yielded 50% coverage. The combination of data covered 88% of the target protein sequence. A plot of sequence coverage versus hydrophobicity shows that peptides from WaLP and MaLP digests are almost solely responsible for coverage of the transmembrane segments (Figure 3.10C).

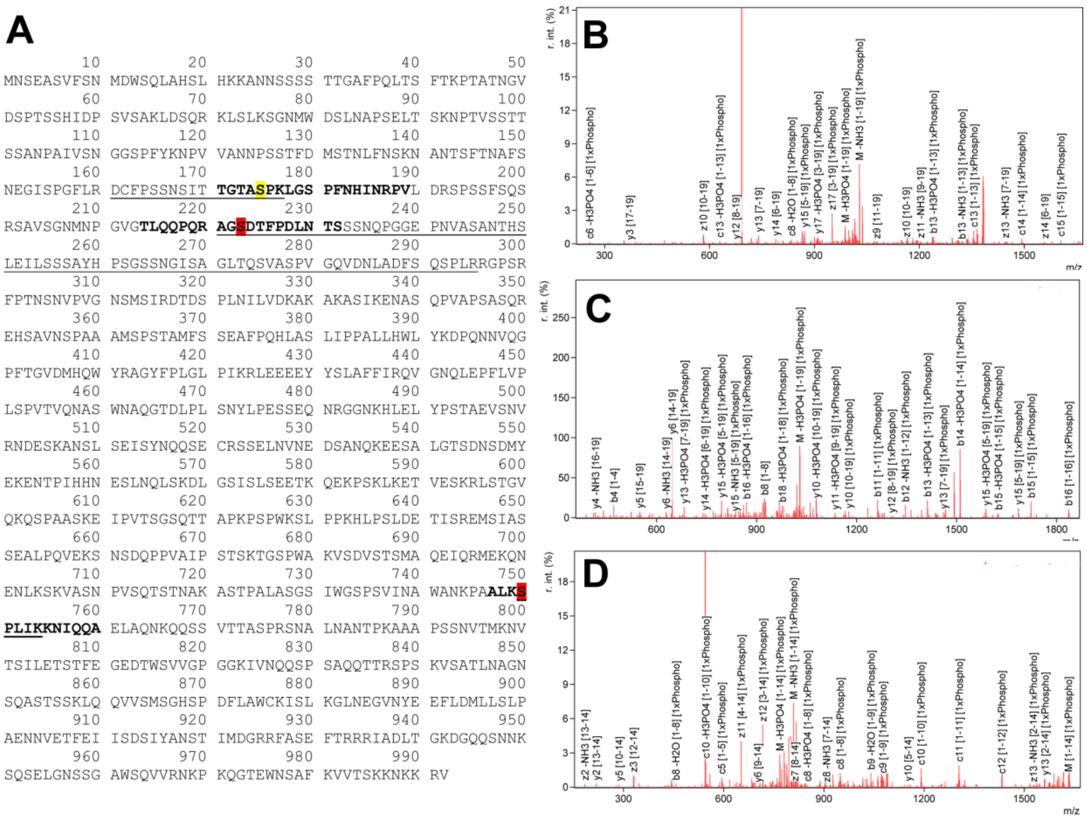


Figure 3.10. (A) Sequence of MPD2 showing S175 (yellow), a previously reported phosphorylation site, S223 and S750 (red), which are phosphorylation sites that haven't been reported before. (B) Annotated spectrum from the +3 charge state precursor of the peptide TGTApSPKLGSPFNHINRPV fragmented by ETD. (C) Annotated spectrum from the +2 charge state precursor of the peptide TLQQPQRAGpSDTFPDLNTS fragmented by CID. (D) Annotated spectrum from the +3 charge state precursor of the peptide ALKpSPLIKKNIQQA fragmented by ETD. The peptide mass information is given in Appendix Table 3.2.

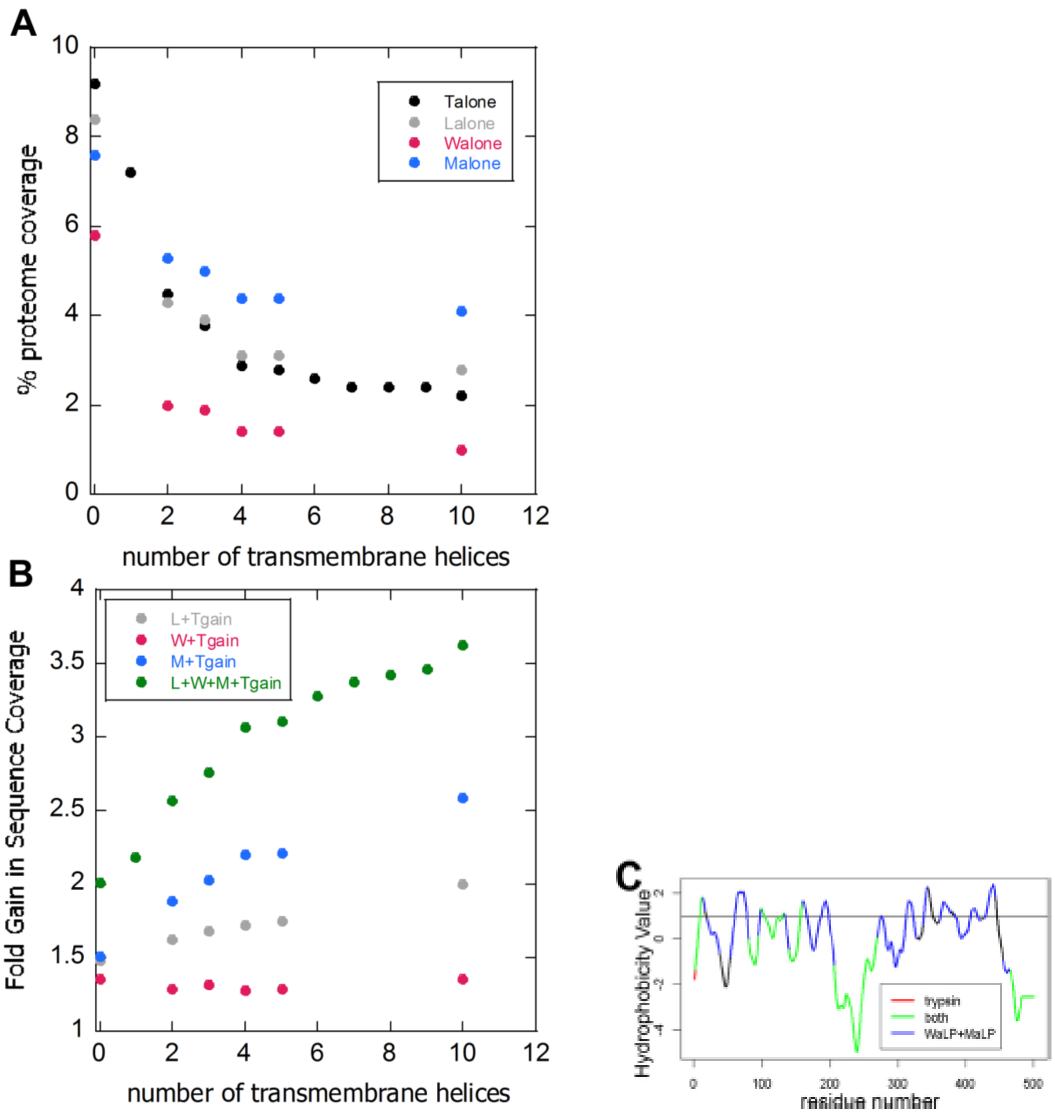


Figure 3.11. MaLP and WaLP improve protein sequence coverage of transmembrane helices. (A) Each protease digest dataset; trypsin (black), LysC (grey), WaLP (red), and MaLP (blue), was evaluated for the amount of protein sequence that was covered in relation to how many transmembrane helices were predicted to be in each protein. The data show that MaLP covers a higher amount of sequence that is predicted to be from proteins containing transmembrane helices. (B) The fold gain of proteome coverage was evaluated for the trypsin dataset combined with Lys C (grey), WaLP (red), MaLP (blue) and all four (green). For the four datasets combined, proteome coverage for proteins with at least four predicted transmembrane helices is increased over three-fold. (C) Plot of the hydrophobicity vs. coverage of the glucose transporter-5 sequence when various combinations of proteases are used for the digestion. WaLP and MaLP cover the more hydrophobic regions whereas trypsin does not.

9. Discussion

The use of alternative proteases has the potential to expand proteome coverage, affording gains in PTM coverage as well as identification of splice variants. In this work, we explore the utility of two proteases that have not been used for proteomics before; WaLP and MaLP. These proteases retain activity in harsh denaturing conditions. They improve coverage of an in-gel digested protein. Combining data from WaLP, MaLP, trypsin and LysC results in nearly 100% coverage of protein sequences in standard mixtures. Thus, WaLP and MaLP digestion will likely prove to be useful for increasing coverage of protein sequences in proteomics, particularly when increased coverage is required for a targeted experiment or when appropriate tryptic cleavage sites are not present.

One possible advantage of WaLP and MaLP are that they cleave at aliphatic residues (A, V, T, S for WaLP; L, F, V for MaLP). Chymotrypsin, which cleaves after aromatic residues (F, Y, W), has also been used to expand protein sequence coverage, but we could not find examples of where chymotrypsin was used in studies of complex proteomes. WaLP and MaLP retain activity throughout long digestion times whereas chymotrypsin does not, potentially making WaLP and MaLP better for improving coverage of proteins in complex proteome mixtures.

The fact that WaLP and MaLP cleave at non-polar residues, however, presented some challenges when using them in global proteomics experiments. The first challenge was their semi-specific substrate specificity. WaLP and MaLP were shown to cleave after several common non-polar residues. Compared with termini from elastase-digestion reported previously, (V [36.5%], I [34.7%], T [30.3%], S [21.4%], L

[19.5%], M [15.7%], and even H [9.1%]) (Wang et al. 2008), WaLP and MaLP (see Results) are more specific. This semi-specificity would be expected to generate many single amino acids and short peptides, which would not be useful in unique sequence determination for proteomics. Surprisingly, the average length of the peptides identified from digests with WaLP were the same as those from trypsin, and peptides from digestion by MaLP were slightly longer than that obtained with trypsin. Figure 3.5 suggests that the substrate recognition preference of WaLP and MaLP extend beyond the P1 position, both before and after the position of cleavage, which is consistent with previous work showing WaLP recognizes at least four amino acids past the position of cleavage (Schellenberger et al. 1994). Thus, WaLP and MaLP target more residues for cleavage, but apparently recognize a longer sequence motif.

Another challenge of the non-polar substrate specificity of WaLP and MaLP is the yield of peptide fragment ions that are useful for sequence determination. WaLP and MaLP peptides yield a significantly lower abundance of y-ions (often scored the highest by database search algorithms). Whereas some 20,000 more MSMS spectra were obtained from the WaLP digest in our ddDT experiment as compared to trypsin, some 2600 fewer peptides were matched to those spectra. This may partly be due to the need to search databases with “no enzyme” specificity. Indeed, searching tryptic digests w/no enzyme for specificity results in 18,520 unique peptide IDs as compared to 21,035. The use of the merged spectra search capability in the MS-GFDB search engine did improve the number of identifications. However, it remains a puzzle as to why the number of peptides identified from the WaLP digest was lower. It is very encouraging that the number of peptides identified from the MaLP digest was significantly higher despite the lower percentage of y-ions. Because WaLP and MaLP don't cleave at K and R, the resultant peptides contain a higher percentage of these positively charged residues. While this doesn't seem to have helped the identification of WaLP peptides, the combination of the higher content of charged residues with the longer average length of the MaLP peptides may have improved the MS2 spectra enough to aid subsequent unambiguous peptide identification.

One striking feature of MaLP specificity is its ability to differentiate I and L, preferring to cleave after L. This observation increases the utility of MaLP digestion, because differences between I/L cannot be resolved by mass alone. Another interesting result was that WaLP and MaLP digests avoid the residue-specific depletion of R and K from trypsin digestion (Fig. 3.8). Thus, WaLP and MaLP are likely to be extremely useful for proteomics analyses of K and/or R- rich sequences.

Sequences identified from WaLP and MaLP digestion are highly complementary to sequences identified from trypsin and LysC digestions. In comparison with Swaney, DL, et al, who doubled proteome coverage relative to trypsin using five separate digestions (Swaney et al. 2010), we achieve double the sequence coverage from only four digestions. The additional coverage is, as expected, beneficial for more comprehensive PTM mapping studies. We show one such example where two new serine phosphorylation sites were identified in MPD2, neither of which is on a peptide that would have been identified from trypsin digestion. Finally, the non-polar substrate specificity of WaLP, but particularly MaLP resulted in a dramatic increase (up to 350%) in proteome coverage of proteins with transmembrane regions. The increase in proteome coverage from all four proteases relative to only trypsin was found to positively correlate with the minimum number of predicted transmembrane helices. Thus, we expect digestion with WaLP and MaLP will find use for comprehensive PTM-mapping studies and especially for a deeper proteomic analysis of membrane proteins.

Chapter III, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in Molecular and Cellular Proteomics. (Meyer, J.G., Kim, S., Maltby, D., Ghassemian, M., Bandeira, N., Komives E.A. (2014). Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. Molecular & Cellular Proteomics, 13, 823-835.)

D. References

Baker, D., J. L. Sohl and D. A. Agard (1992). "A protein-folding reaction under kinetic control." Nature**356**: 263-265.

Bone, R., J. L. Silen and D. A. Agard (1989). "Structural plasticity broadens the specificity of an engineered protease." Nature**339**(6221): 191-195.

Chalkley, R. J. and K. R. Clauser (2012). "Modification site localization scoring: strategies and performance." Mol Cell Proteomics**11**(5): 3-14.

Colaert, N., K. Helsens, L. Martens, J. Vandekerckhove and K. Gevaert (2009). "Improved visualization of protein consensus sequences by iceLogo." Nature Methods**6**(11): 786-787.

de Godoy, L. M. F., J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther and M. Mann (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature**455**(7217): 1251-1254.

Elias, J. E., F. D. Gibbons, O. D. King, F. P. Roth and S. P. Gygi (2004). "Intensity-based protein identification by machine learning from a library of tandem mass spectra." Nature Biotech**22**(2): 214-219.

Graham, L. D., K. D. Haggett, P. A. Jennings, D. S. Le Brocque, R. G. Whittaker and P. A. Schober (1993). "Random mutagenesis of the substrate-binding site of a serine protease can generate enzymes with increased activities and altered primary specificities." Biochemistry**32**(24): 6250–6258.

Gupta, N., N. Bandeira, U. Keich and P. A. Pevzner (2011). "Target-decoy approach and false discover rate: When things may go wrong." J Am Soc Mass Spectrom**22**(7): 1111-1120.

Guthals, A. and N. Bandeira (2012). "Peptide identification by tandem mass spectrometry with alternate fragmentation modes." Mol Cell Proteomics**11**(9): 550-557.

Guthals, A., K. Klauser, A. M. Frank and N. Bandeira (2013). "Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides." J Proteom Res**12**: 2846−2857.

Huang, Y., J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith and V. H. Wysocki (2005). "Statistical Characterization of the Charge State and Residue Dependence of Low-Energy CID Peptide Dissociation Patterns." Anal Chem**77**(18): 5800-5813.

Jaswal, S. S., J. L. Sohl, J. H. Davis and D. A. Agard (2002). "Energetic landscape of [alpha]-lytic protease optimizes longevity through kinetic stability." Nature**415**(6869): 343–346.

Keller, A., J. Eng, N. Zhang, X. Li and R. Aebersold (2005). "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." Mol Syst Biol: 1:2005.0017.

Keller, A., A. I. Nesvizhskii, E. Kolker and R. Aebersold (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Anal Chem**74**(20): 5383–5392.

Kessner, D., M. Chambers, R. Burke, D. Agus and P. Mallick (2008). "ProteoWizard: open source software for rapid proteomics tools development." Bioinformatics**24**(21): 2534-2536.

Kim, S., N. Gupta and P. P.A. (2008). "Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases." J Proteome Res.**7**(8): 3354-3363.

Kim, S., N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck and P. A. Pevzner (2010). "The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search." Mol Cell Proteom**9**(12): 2840-2852.

Krogh, A., B. Larsson, G. von Heijne and E. L. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a Hidden Markov Model: Application to complete genomes." J Mol Biol**305**(3): 567–580.

MacCoss, M. J., W. H. McDonald, A. Saraf, R. Sadygov, J. M. Clark, J. J. Tasto, K. L. Gould, D. Wolters, M. Washburn, A. Weiss, J. I. Clark and J. R. Yates (2002). "Shotgun identification of protein modifications from protein complexes and lens tissue." Proc Nat Acad Sci USA**99**(12): 7900-7905.

Mace, J. E. and D. A. Agard (1995). "Kinetic and structural characterization of mutations of glycine 216 in alpha-lytic protease: a new target for engineering substrate specificity." J Mol Biol**254**: 720-736.

Mace, J. E., B. J. Wilk and D. A. Agard (1995). "Functional linkage between the active site of α-lytic protease and distant regions of structure: Scanning alanine mutagenesis of a durface loop sffects sctivity and dubstrate dpecificity." J Mol Biol**251**(1): 116-134.

Masuda, T., N. Saito, M. Tomita and Y. Ishihama (2009). "Unbiased quantitation of Escherichia coli membrane proteome using phase transfer surfactants." Mol Cell Proteomics**8**(12): 2770-2777.

Masuda, T., M. Tomita and Y. Ishihama (2008). "Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis." J Proteom Res**7**(2): 731-740.

Meyer, J. G. and E. A. Komives (2012). "Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry." J Am Soc Mass Spectrom**23**(8): 1390-1399.

Michalski, A., N. Neuhauser, J. Cox and M. Mann (2012). "A systematic investigation into the nature of tryptic HCD spectra." J Proteome Res**11**(11): 5479-5491.

Nagaraj, N., N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm and M. Mann (2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap." Mol Cell Proteomics**11**(3): M111.013722.

Nesvizhskii, A. I., A. Keller, E. Kolker and R. Aebersold (2003). "A statistical model for identifying proteins by tandem mass spectrometry." Anal Chem**75**(17): 4646-4658.

Olsen, J. V., J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov and O. Lange (2009). "A Dual pressure linear ion trap Orbitrap instrument with very high sequencing speed." Mol Cell Proteomics**8**(12): 2759 –2769.

Rietschel, B., T. N. Arrey, B. Meyer, S. Bornemann, M. Schuerken, M. Karas and A. Poetsch (2009). "Elastase digests: new ammunition for shotgun membrane proteomics." Mol Cell Proteomics**8**(5): 1029-1043.

Schellenberger, V., C. W. Turck and W. J. Rutter (1994). "Role of the S’ subsites in serine protease catalysis. Active-site mapping of rat chymotrypsin, rat trypsin, alpha-lytic protease, and cercarial protease from Schistosoma mansoni." Biochemistry**33**(14): 4251-4257.

Shen, Y., N. Tolić, F. Xie, R. Zhao, S. O. Purvine, A. A. Schepmoes, R. J. Moore, G. A. Anderson and R. D. Smith (2011). "Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: Comparison of peptide identification methods." J Proteom Res**10**(9): 3929-3943.

Silen, J. L., D. Frank, A. Fujishige, R. Bone and D. A. Agard (1989). "Analysis of prepro-alpha-lytic protease expression in Escherichia coli reveals that the pro region is required for activity." J Bacteriol**171**(3): 1320 -1325.

Sleno, L. and D. A. Volmer (2004). "Ion activation methods for tandem mass spectrometry." J Mass Spectrom**39**(10): 1091-1112.

Sohl, J. L., S. S. Jaswal and D. A. Agard (1998). "Unfolded conformations of [alpha]-lytic protease are more stable than its native state." Nature**395**(6704): 817-819.

Swaney, D. L., G. C. McAlister and J. J. Coon (2008). "Decision tree-driven tandem mass spectrometry for shotgun proteomics." Nat Meth**5**(11): 959-964.

Swaney, D. L., C. D. Wenger and J. J. Coon (2010). "Value of using multiple proteases for large-scale mass spectrometry-based proteomics." J Proteom Res**9**(3): 1323-1329.

Syka, J. E. P., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." Proc Nat Acad Sci USA**101**(26): 9528 -9533.

Tabb, D. L., Y. Huang, V. H. Wysocki and J. R. Yates (2004). "Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides." Anal Chem**76**(5): 1243-1248.

Tabb, D. L., L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin and J. R. Yates (2003). "Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides." Anal Chem**75**(5): 1155-1163.

Team, R. D. C. (2011). "R: A Language and Environment for Statistical Computing."

von der Haar, T. (2007). "Optimized protein extraction for quantitative proteomics of yeasts." PLoS ONE**2**(10): e1078.

Walther, T. C. and M. Mann (2010). "Mass spectrometry-based proteomics in cell biology." J Cell Biol**190**(4): 491-500.

Wang, B., R. Malik, E. A. Nigg, R. Körner and E. o. t. L.-S. P. E. f. L.-S. P. A. A. C. Dec;80(24):9526–9533. (2008). "Evaluation of the low-specificity protease elastase for large-scale phosphoproteome analysis." Anal Chem**80**(24): 9526-9533.

Wang, Y., Y. Feng, M. A. Gritsenko, Y. Wang, T. Clauss, T. Liu, Y. Shen, M. E. Monroe, D. Lopez-Ferrer, T. Reno, R. J. Moore, R. L. Klemke, D. G. Camp II and R. D. Smith (2011). "Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A vells." Proteomics**11**(10): 2019-2026.

Wiśniewski, J. R., K. Duś and M. Mann (2012). "Proteomic workflow for analysis of archival formalin fixed and paraffin embedded clinical samples to a depth of 10,000 proteins." PROTEOMICS – Clinical Applications**7**(3-4): 225-233.

Wiśniewski, J. R. and M. Mann (2012). "Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis." Anal Chem**84**(6): 2631–2637.

Wysocki, V. H., G. Tsaprailis, L. L. Smith and L. A. Breci (2000). "Mobile and localized protons: a framework for understanding peptide dissociation." J Mass Spectrom**35**: 1399-1406.

Chapter IV

In silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage

A. Introduction

In the post genome era, biologists have sought system-wide measurements of RNA, proteins, and metabolites, termed transcriptomics, proteomics, and metabolomics, respectively. Shotgun, or bottom-up, proteomics has become the most comprehensive method for proteome identification and quantification (Zhang et al. 2013). However, observed protein sequence coverage is often low. The ability to cover 100% of protein sequences in a biological system was likened to surrealism in a recent review by Karas et al (B. Meyer, et al. 2010). Multiple steps in the traditional shotgun proteomics workflow contribute to the deficit in observed sequence coverage, including: proteome isolation, proteome digestion, peptide separation, peptide MS/MS, and identification by peptide-spectrum matching. Proteome isolation has been extensively evaluated (Gilmore and Washburn 2010; Rey et al. 2010). Several types of peptide separation have been have been explored (Motoyama and Yates 2008; Wang et al. 2011,Betancourt et al. 2013). Mass spectrometers are becoming more sensitive and versatile (Michalski et al. 2011; Olsen et al. 2009; Frese et al. 2011). Peptide-spectrum matching algorithms are adapting to new data types (Chalkley et al. 2008) and becoming more sensitive (Shen et al. 2011; Kim et al. 2010). Proteome fragmentation into sequence-able peptides is one step with significant room for improvement. DNA sequencing relies on sequence fragmentation into readable pieces by mechanical force (Linnarsson 2010), which produces a nearly uniform distribution of fragment lengths. In comparison, proteome fragmentation is generally accomplished by targeting one or more amino acid residues for cleavage, and therefore, the protein cleavage can be likened to a Poisson process that produces an exponential distribution of peptide lengths.

Numerous papers have described the application of new digestion strategies for proteome analysis (Rietschel et al. 2009; Choudhary et al. 2002; Moura et al. 2013; Tran et al. 2010), however, no single strategy has emerged as optimal. The greatest observed proteome coverage has plateaued around 25%. 24.6% of the human proteome was recently observed (Neuhauser et al. 2013), but this was obtained from over 1,000 MS/MS data files that allowed identification of over 260,000 peptide sequences using a new high performance data analysis package. Similarly, multiple protease digests of yeast resulted in 25.2% coverage (Swaney, Wenger, and Coon 2010). Therefore, improved strategies for proteome digestion are needed to allow observation of a complete proteome.

An innovative example demonstrating the application of multiple enzyme digestion (MED) was recently published by Wiśniewski and Mann (Wiśniewski and Mann 2012), which demonstrated the utility of multi-enzyme digestion coupled to filter-aided sample preparation (Wisniewski et al. 2009) (MED-FASP, Figure 4.1). This work extends a previous work that described size exclusion to isolate long tryptic peptides for additional digestion (Tran et al. 2010). Wiśniewski and Mann compared gains afforded by iterative digestion using various proteases (i.e. GluC, ArgC, LysC, or AspN) followed by trypsin. Their work concluded that iterative digestion with LysC followed by Trypsin allowed 31% more protein identifications and a 2-fold gain in observed phosphopeptides for a particular protein. Their work led me to optimize

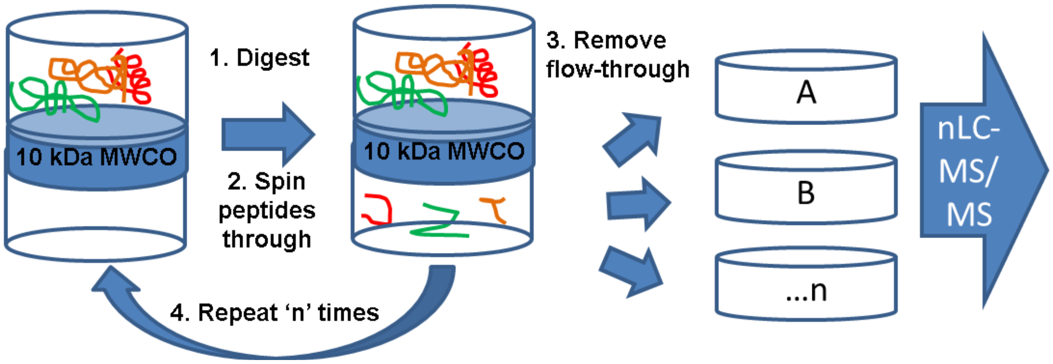


Figure 4.1. A cartoon descriping the MED-FASP digestion strategy. Proteins retained on an ultrafilter are digested and then peptides are spun through the size-based filter. Large, undigested protein sequences retained above the ultrafilter are then digested again. This sequence of digestion and elution can be repeated with an arbitrary number of digestions.

iterative digestion *in silico* with the hope of identifying a testable digestion strategy that can theoretically achieve complete proteome coverage.

B. Materials and Methods

1. *In silico* proteome digestion

The S. cerevisiae proteome file in FASTA format was downloaded from uniprot on June 20th, 2012. Proteome digestion simulations were accomplished using scripts written in [R] (R Development Core Team 2008). Considered protease specificities include c-terminal of: R/K (trypsin), L (LeuC theoretical cleavage agent), E (GluC), and K (LysC). Additionally, simulations utilized chemical digestion agents (Crimmins, et al. 2001), including cyanogen bromide (CNBr) (Kaiser and Metzka 1999; Andreev et al. 2010) for cleavage C-terminal of M, 3-Bromo-3-methyl-2-(2-nitrophenylthio)-3H-indole (BNPS-skatole) for cleavage C-terminal of W (Vestling et al. 1994), and 2-nitro-5-thiocyanobenzoic acid (NTCB) for cleavage N-terminal of C (Jacobson et al. 1973; Iwasaki et al. 2009). Peptide populations were filtered using both length and molecular weight constraints. Since the filtration thresholds affect the proteome coverage prediction, multiple cutoff values are compared. The [R] code is available at: https://github.com/jgmeyerucsd/ProteomeDigestSim.

C. RESULTS

1. Minimum unique peptide length

The probability of a sequence being unique can be calculated assuming a random distribution of sequences in the library. The number of sequences of length n can be described by: 20^n. Therefore, any given sequence of length five is likely to occur once in a library of 3,200,000 random amino acid sequences (roughly the number of amino acids in the *S. cerevisiae* proteome). As the number of amino acids in the database grows, a peptide sequence must be longer to expect uniqueness. The human proteome contains 11,323,900 amino acids (not including isoforms, downloaded from uniprot on October 22nd, 2013), and therefore, for a sequence to be unique, it must be at least of length six. Of course, due to common sequence motifs there are less unique peptide sequences in a proteome than would be found in a random library.

2. Peptide length distributions from various cleavages

Initial *in silico* digestions using single cleavage agents were used to compare the resulting peptide lengths (Figure 4.2). Many peptide sequences are too short to uniquely match a protein. For all digestion agents, the most frequent peptide length produced is one. Generation of a single amino acid would arise when the target residue is next to itself in the protein. Notably, over 25% of theoretical peptides from trypsin digestion, which cleaves after 11.7% of all residues, are of length one. Not surprisingly, the observable proportion of the residue targeted for cleavage correlates with the resulting average peptide length (Figure 4.3); more common cleavage targets produce shorter average peptide lengths. Additionally, the residue-level coverage was found to depend on digestion. Proteome cleavage after more common residues results in depletion of the target residues (Figure 4.4), which is expected to result from production of peptides that are too short to uniquely match a protein sequence. However, cleavage after rare residues results in enriched coverage of the target

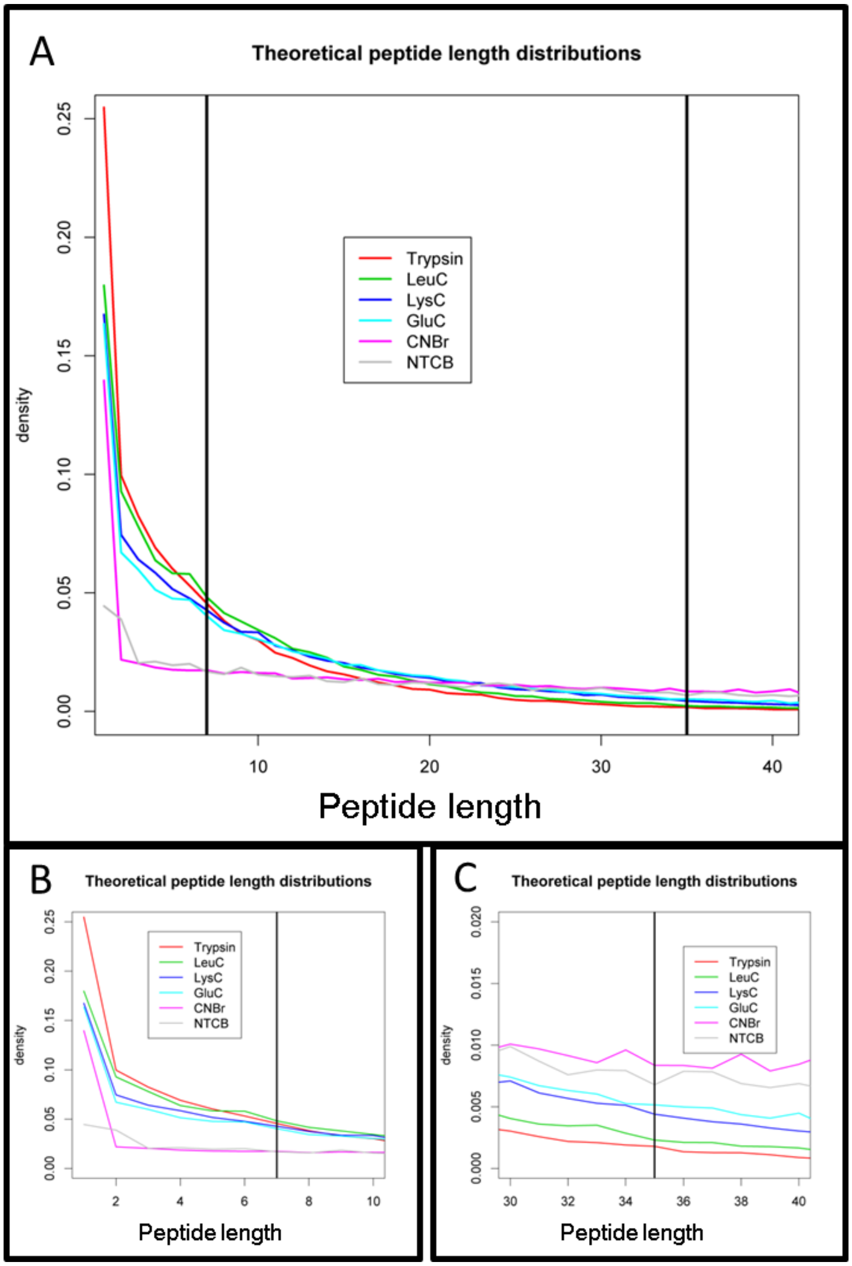
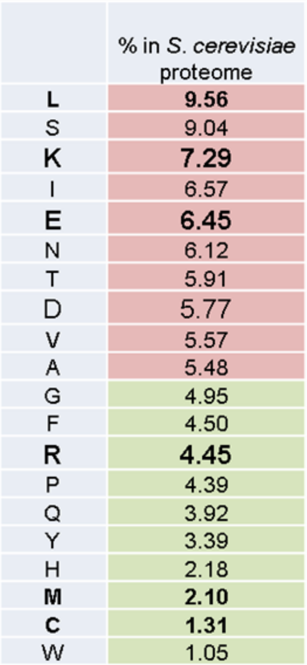


Figure 5.2. Theoretical peptide length distributions produced from various cleavage agents. (A) Size frequency distributions (density) of peptides from proteome digestion by five real (i.e. trypsin, LysC, GluC, CNBr, NTCB) and one theoretical cleavage agent (LeuC). The vertical black lines at 7 and 35 indicate general peptide identification size limits.(B) The same distribution focused on the region from 1-10 amino acids. (C) The view focused on the region between 30-40 amino acids.



	% in *S. cerevisiae* proteome
L	**9.56**
S	9.04
K	**7.29**
I	6.57
E	**6.45**
N	6.12
T	5.91
D	5.77
V	5.57
A	5.48
G	4.95
F	4.50
R	**4.45**
P	4.39
Q	3.92
Y	3.39
H	2.18
M	**2.10**
C	**1.31**
W	1.05

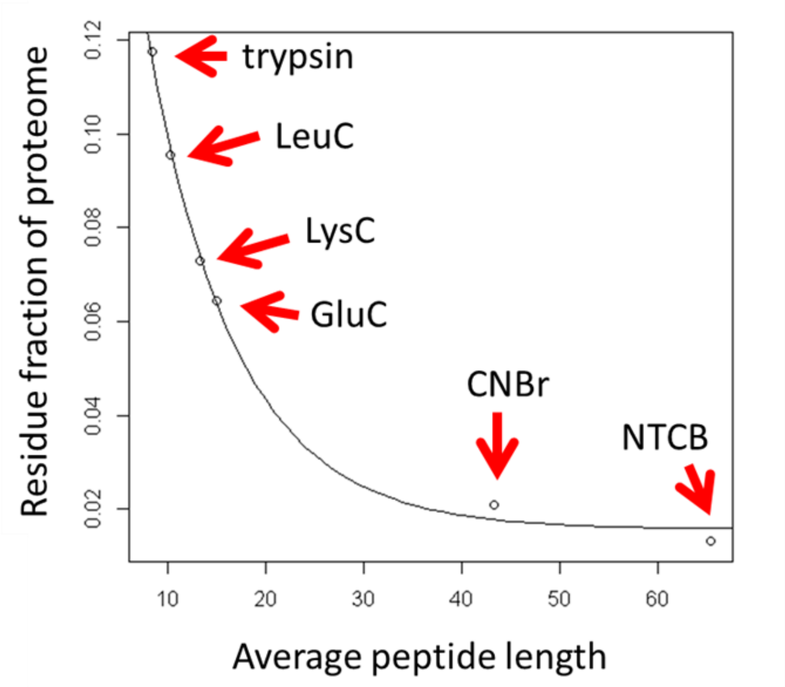


Figure 4.3. Correlation between abundance of the residue targeted for cleavage and the resulting average peptide length. Proteome cleavage targeting abundant residues result in lower average peptide lengths; proteome cleavage targeting rare residues results in higher average peptide length. The line shows the data fit to an exponential equation.

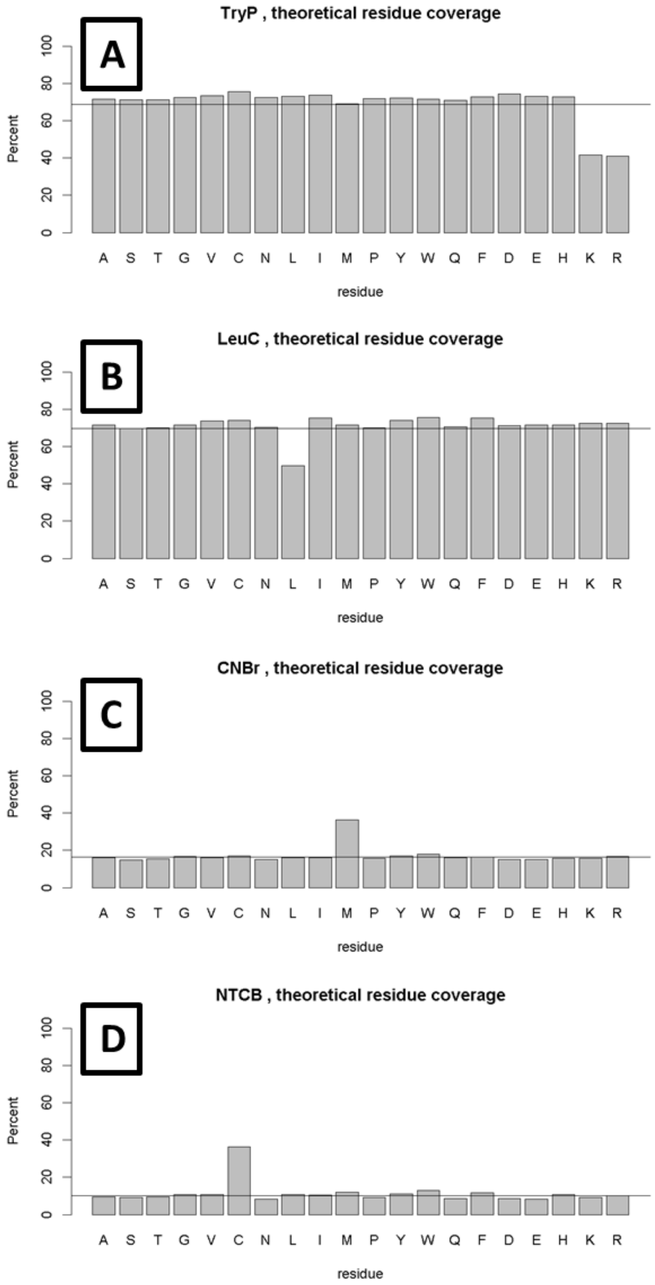


Figure 4.4. Residue-level coverage observed for various cleavage agents. Proteome cleavage of more common amino acids, such as with (A) trypsin or the theoretical cleavage after (B) Leucine, result in residue-specific depletion of the target residues. However, cleavage of rare amino acids, such as (C) Methionine or (D) Cysteine, results in residue-specific enrichment of the target residues.

residue. This result was also observed by amino acid analysis of proteome digestions in recent work (J. G. Meyer et al. 2014).

3. Comparison of peptide filtration parameters

The theoretical distribution of peptides passing through a MWCO ultrafilter certainly does not match the actual distribution. Denatured peptides and proteins are effectively larger than folded proteins, and in fact, it was found that even 30kDa or 50kDa cut-off ultrafilters perform better for peptide yield than 10 kDa cut-off ultrafilters (Wisniewski, et al. 2011), despite the inability to identify such large peptide sequences by bottom-up proteomics. Therefore, multiple length constraints were compared for their influence on the predicted proteome coverage. Figure 4.5 shows how various minimum peptide length values affect residue-level depletion and theoretical proteome coverage. As the minimum length increases, total coverage decreases and depletion of R/K increases. Figure 4.6 shows how different upper length thresholds change theoretical coverage. Intuitively, raising the upper length limit of identifiable peptides increases total predicted proteome coverage. Interestingly, although total predicted coverage increases, the coverage of R/K stays around 60%. Since peptide MW also determines identifiable peptides, and peptides above 5kDa are unlikely to be identified with current MSMS technology, an upper limit of 5 kDa was used for subsequent digest simulations. A lower length limit of 7 amino acids was used because this length is more likely to be relevant to actual proteomics experiments.

4. Comparison of digestion iterations

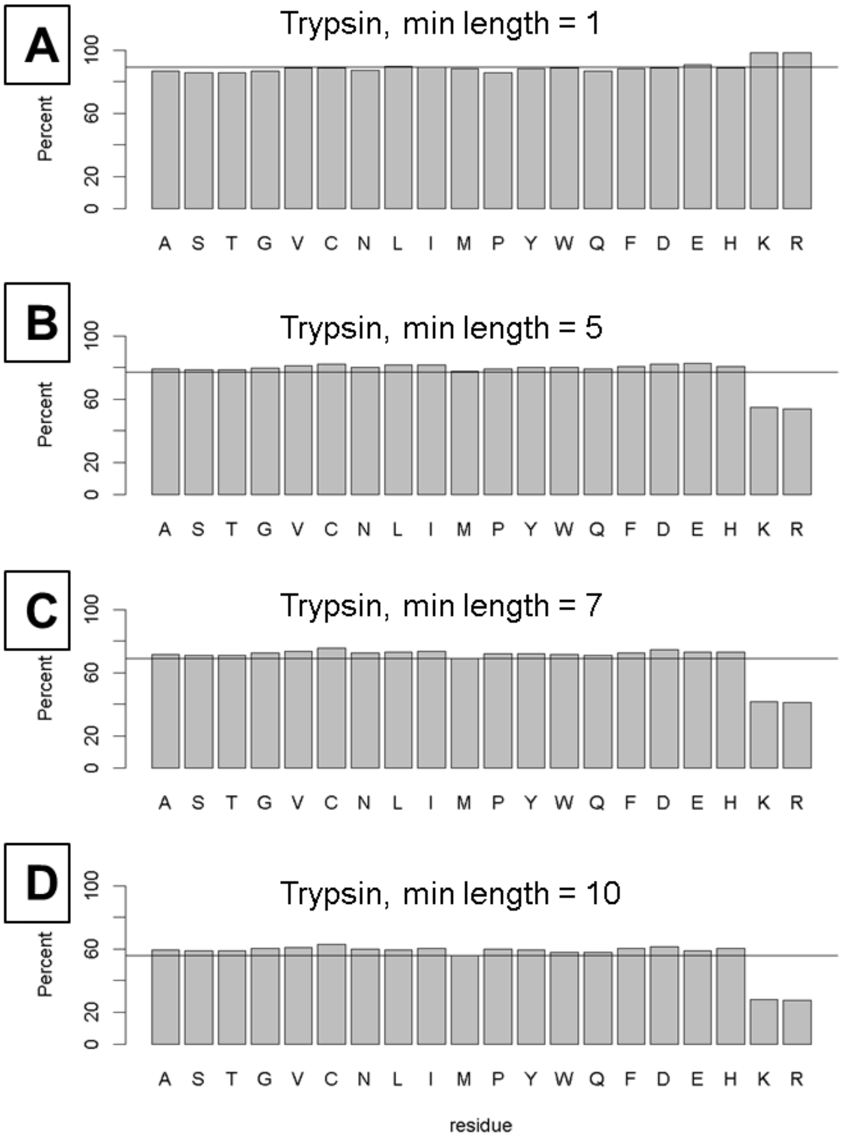


Figure 4.5. Effect of minimum peptide length on proteome coverage and residue-level depletion. Residue-level coverage predicted after trypsin digestion keeping all peptides with lengths between: (A) 1-35, (B) 5-35, (C) 7-35, and (D) 10-35.

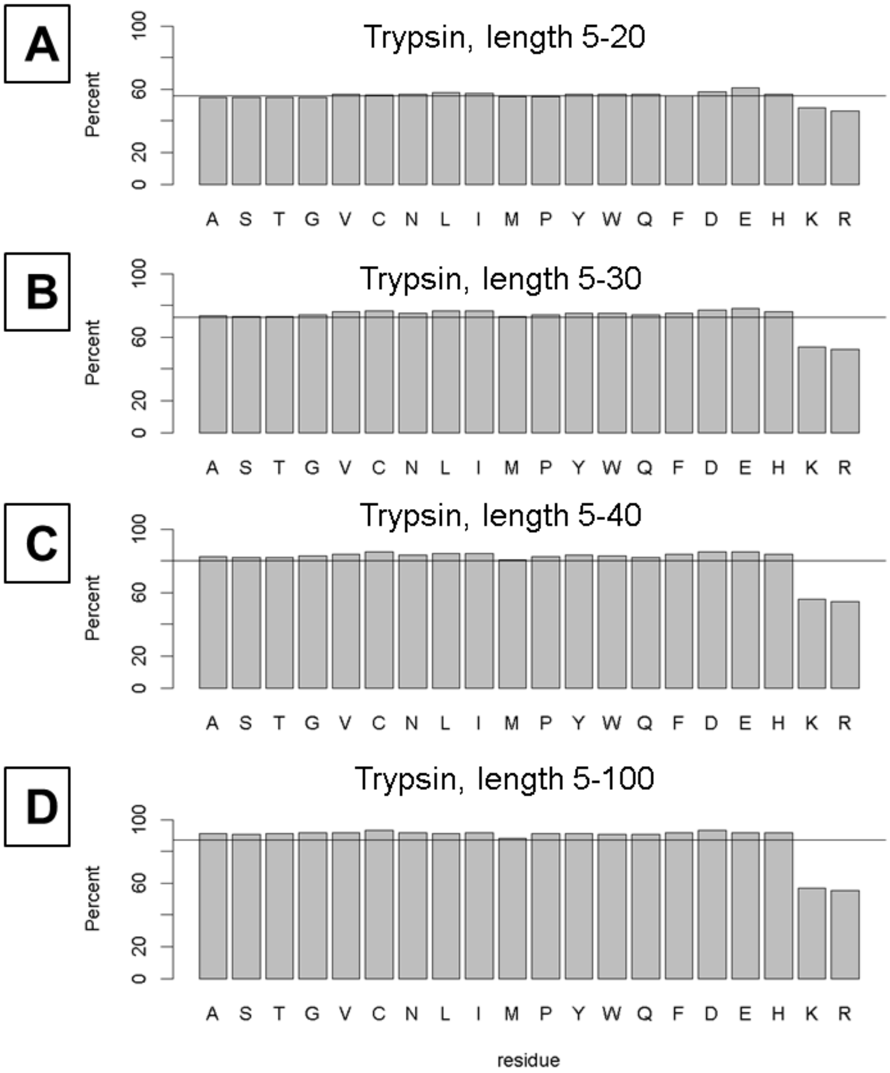


Figure 4.6. Effect of upper length limit on predicted proteome coverage. Theoretical coverage keeping peptides with length (A) 5-20, (B) 5-30, (C) 5-40, and (D) 5-100 residues. As the upper length limit increases, the theoretical coverage maximum increases.

Digest simulations for various digestion iterations were performed to compute theoretical proteome coverage for various iterative digestions. Simulations confirm that iterative digestion offers theoretically greater coverage of the proteome when the sequence of digestions starts with the protease targeting the rarest residue first (Table 4.1). As expected, reversal of the optimal digestion sequence results in a negligible improvement to proteome coverage as compared to the limit from using trypsin digestion alone.

5. Proposed iterative digestion strategy

An ideal iterative cleavage strategy must limit sample processing steps, and must take place under conditions that are compatible with the ultrafiltration device. Further, because tryptophan fluorescence can be used to quantify peptide yield from each digestion, chemical cleavage after tryptophan should initially be omitted since it destroys the fluorophore. Therefore, an ultrafilter-compatible strategy, with a balance between sample processing and predicted gains in coverage, is the sequence: NTCB, CNBr, LysC, and Trypsin. Implementation of this method will likely require optimization at various steps.

6. Conclusions

This work provides a publically accessible computational framework for simulation of iterative proteome digestion that can be used with any input protein sequence database to optimize proteome coverage. Further, this works demonstrates how the choice of proteome digestion agent affects the predicted proteome coverage due to the distribution of peptide lengths that are produced. This work also shows

Table 4.1. Theoretical upper limits of coverage upon digestion with various cleavage agents using the MED-FASP strategy. Iterative cleavage of the proteome starting with the rarest amino acids first results in the greatest theoretical proteome coverage of 91.1%. The last row gives the theoretical limit from digestion in the reversed sequence of cleavage, which provides minimal improvement to theoretical proteome coverage. Peptides were filtered after each digest keeping those with MW >5 kDa for additional digestion.

Digestion strategy	Theoretical coverage limit (%)
Trypsin	69.5
LysC	67.1
GluC	62.7
AspN	63.1
ArgC	52.4
CNBr	22.4
NTCB	13.6
TrpC	10.9
LysC, Trypsin	81.2
GluC, Trypsin	81.1
CNBr, LysC, Trypsin	84.4
NTCB, CNBr, LysC, Trypsin	86.3
TrpC, NTCB, CNBr, ArgC, GluC, Trypsin	87.9
TrpC, NTCB, CNBr, ArgC, AspN, GluC, Trypsin	91.1
Trypsin, GluC, AspN, ArgC, CNBr, NTCB, TrpC	74.2

how various digestion agents affect proteome coverage at the residue level. Proteome cleavage targeting common residues results in depletion of the cleaved residue, but proteome cleavage after rare residues results in enrichment of the target residue. Finally, this paper finds that the best theoretical proteome coverage is achieved by an iterative digestion strategy that limits production of short peptides by cleaving the rarest residues first.

Chapter IV, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in ISRN Computational Biology. (Meyer, J.G. (2014). In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. ISRN Computational Biology, 2014, 1-7.)

D. References

Andreev, Yaroslav A., Sergey A. Kozlov, Alexander A. Vassilevski, and Eugene V. Grishin. 2010. “Cyanogen Bromide Cleavage of Proteins in Salt and Buffer Solutions.” Anal Bioc 407 (1): 144–46. doi:10.1016/j.ab.2010.07.023.

Betancourt, Lázaro H., Pieter-Jan De Bock, An Staes, Evy Timmerman, Yasset Perez-Riverol, Aniel Sanchez, Vladimir Besada, Luis Javier Gonzalez, Joël Vandekerckhove, and Kris Gevaert. 2013. “SCX Charge State Selective Separation of Tryptic Peptides Combined with 2D-RP-HPLC Allows for Detailed Proteome Mapping.” J Prot 91 (0): 164–71. doi:10.1016/j.jprot.2013.06.033.

Chalkley, Robert J., Peter R. Baker, Katalin F. Medzihradszky, Aenoch J. Lynn, and A. L. Burlingame. 2008. “In-Depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types.” Mol Cell Proteomics 7 (12): 2386–98. doi:10.1074/mcp.M800021-MCP200.

Choudhary, Gargi, Shiaw-Lin Wu, Paul Shieh, and William S. Hancock. 2002. “Multiple Enzymatic Digestion for Enhanced Sequence Coverage of Proteins in Complex Proteomic Mixtures Using Capillary LC with Ion Trap MS/MS.” J Proteome Res 2 (1): 59–67. doi:10.1021/pr025557n.

Crimmins, Dan L., Sheenah M. Mische, and Nancy D. Denslow. 2001. “Chemical Cleavage of Proteins in Solution.” Current Protocols in Protein Science. John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471140864.ps1104s40.

Frese, Christian K., A. F. Maarten Altelaar, Marco L. Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert J. R. Heck, and Shabaz Mohammed. 2011. “Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos.” J Proteome Res 10 (5): 2377–88. doi:10.1021/pr1011729.

Gilmore, Joshua M., and Michael P. Washburn. 2010. “Advances in Shotgun Proteomics and the Analysis of Membrane Proteomes.” Model Organism Proteomics 73 (11): 2078–91. doi:10.1016/j.jprot.2010.08.005.

Iwasaki, Mio, Takeshi Masuda, Masaru Tomita, and Yasushi Ishihama. 2009. “Chemical Cleavage-Assisted Tryptic Digestion for Membrane Proteome Analysis.” J Proteome Res 8 (6): 3169–75. doi:10.1021/pr900074n.

Jacobson, Gary R., Martin H. Schaffer, George R. Stark, and Thomas C. Vanaman. 1973. “Specific Chemical Cleavage in High Yield at the Amino Peptide Bonds of Cysteine and Cystine Residues.” J Biol Chem 248 (19): 6583–91.

Kaiser, Raymond, and Lorraine Metzka. 1999. “Enhancement of Cyanogen Bromide Cleavage Yields for Methionyl-Serine and Methionyl-Threonine Peptide Bonds.” Anal Bioc 266 (1): 1–8. doi:10.1006/abio.1998.2945.

Kim, Sangtae, Nikolai Mischerikow, Nuno Bandeira, J. Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J. R. Heck, and Pavel A. Pevzner. 2010. “The Generating Function of CID, ETD and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search.” Mol Cell Proteomics, doi:10.1074/mcp.M110.003731.

Linnarsson, Sten. 2010. “Recent Advances in DNA Sequencing Methods – General Principles of Sample Preparation.” Special Issue Celebrating the 60-Year Anniversary of ECR and the 200-Year Anniversary of the Karolinska Institute 316 (8): 1339–43. doi:10.1016/j.yexcr.2010.02.036.

Meyer, Bjoern, Dimitrios G. Papasotiriou, and Michael Karas. 2010. “100% Protein Sequence Coverage: A Modern Form of Surrealism in Proteomics.” Amino Acids 41 (2): 291–310. doi:10.1007/s00726-010-0680-6.

Meyer, Jesse G., Sangtae Kim, David Maltby, Majid Ghassemian, Nuno Bandeira, and Elizabeth A. Komives. 2014. “Expanding Proteome Coverage with Orthogonal-Specificity Alpha-Lytic Proteases.” Mol Cell Proteomics, doi:10.1074/mcp.M113.034710.

Michalski, Annette, Eugen Damoc, Jan-Peter Hauschild, Oliver Lange, Andreas Wieghaus, Alexander Makarov, Nagarjuna Nagaraj, Juergen Cox, Matthias Mann, and Stevan Horning. 2011. “Mass Spectrometry-Based Proteomics Using Q Exactive, a High-Performance Benchtop Quadrupole Orbitrap Mass Spectrometer.” Mol Cell Proteomics, 10 (9). doi:10.1074/mcp.M111.011015.

Motoyama, Akira, and John R. Yates. 2008. “Multidimensional LC Separations in Shotgun Proteomics.” Anal Chem 80 (19): 7187–93. doi:10.1021/ac8013669.

Moura, Hercules, Rebecca R. Terilli, Adrian R. Woolfitt, Yulanda M. Williamson, Glauber Wagner, Thomas A. Blake, Maria I. Solano, and John R. Barr. 2013. “Proteomic Analysis and Label-Free Quantification of the Large Clostridium Difficile Toxins.” Int J Proteomics 2013: 1–10. doi:10.1155/2013/293782.

Neuhauser, Nadin, Nagarjuna Nagaraj, Peter McHardy, Sara Zanivan, Richard Scheltema, Jürgen Cox, and Matthias Mann. 2013. “High Performance Computational Analysis of Large-Scale Proteome Data Sets to Assess Incremental Contribution to Coverage of the Human Genome.” J Proteome Res 12 (6): 2858–68. doi:10.1021/pr400181q.

Olsen, Jesper V., Jae C. Schwartz, Jens Griep-Raming, Michael L. Nielsen, Eugen Damoc, Eduard Denisov, Oliver Lange, et al. 2009. “A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed.” Mol Cell Proteomics, 8 (12): 2759–69. doi:10.1074/mcp.M900375-MCP200.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rey, Martial, Hynek Mrázek, Petr Pompach, Petr Novák, Ludovic Pelosi, Gérard Brandolin, Eric Forest, Vladimír Havlíček, and Petr Man. 2010. “Effective Removal of Nonionic Detergents in Protein Mass Spectrometry, Hydrogen/Deuterium Exchange, and Proteomics.” Anal Chem, 82 (12): 5107–16. doi:10.1021/ac100171m.

Rietschel, Benjamin, Tabiwang N. Arrey, Bjoern Meyer, Sandra Bornemann, Malte Schuerken, Michael Karas, and Ansgar Poetsch. 2009. “Elastase Digests.” Mol Cell Proteomics, 8 (5): 1029–43. doi:10.1074/mcp.M800223-MCP200.

Shen, Yufeng, Nikola Tolić, Samuel O. Purvine, and Richard D. Smith. 2011. “Improving Collision Induced Dissociation (CID), High Energy Collision Dissociation (HCD), and Electron Transfer Dissociation (ETD) Fourier Transform MS/MS Degradome–Peptidome Identifications Using High Accuracy Mass Information.” J Proteome Res, 11 (2): 668–77. doi:10.1021/pr200597j.

Swaney, Danielle L., Craig D. Wenger, and Joshua J. Coon. 2010. “Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics.” J Proteome Res, 9 (3): 1323–29. doi:doi: 10.1021/pr900863u.

Tran, Bao Quoc, Celine Hernandez, Patrice Waridel, Alexandra Potts, Jachen Barblan, Frederique Lisacek, and Manfredo Quadroni. 2010. “Addressing Trypsin Bias in Large Scale (Phospho)proteome Analysis by Size Exclusion Chromatography and Secondary Digestion of Large Post-Trypsin Peptides.” J Proteome Res, 10 (2): 800–811. doi:10.1021/pr100951t.

Vestling, Martha M., Michele A. Kelly, Catherine Fenselau, and Catherine E. Costello. 1994. “Optimization by Mass Spectrometry of a Tryptophan-Specific Protein Cleavage Reaction.” Rapid Comm Mass Spectrometry, 8 (9): 786–90. doi:10.1002/rcm.1290080925.

Wang, Yuexi, Feng Yang, Marina A. Gritsenko, Yingchun Wang, Therese Clauss, Tao Liu, Yufeng Shen, et al. 2011. “Reversed-Phase Chromatography with

Multiple Fraction Concatenation Strategy for Proteome Profiling of Human MCF10A Cells.” Proteomics, 11 (10): 2019–26. doi:10.1002/pmic.201000722.

Wiśniewski, Jacek R., and Matthias Mann. 2012. “Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis.” Anal Chem, 84 (6): 2631–37. doi:10.1021/ac300006b.

Wisniewski, Jacek R., Dorota F. Zielinska, and Matthias Mann. 2011. “Comparison of Ultrafiltration Units for Proteomic and N-Glycoproteomic Analysis by the Filter-Aided Sample Preparation Method.” Anal Bioc, 410 (2): 307–9. doi:doi: 10.1016/j.ab.2010.12.004.

Wisniewski, Jacek R, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. 2009. “Universal Sample Preparation Method for Proteome Analysis.” Nat Meth, 6 (5): 359–62. doi:10.1038/nmeth.1322.

Zhang, Yaoyang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates. 2013. “Protein Analysis by Shotgun/Bottom-up Proteomics.” Chem Rev, 113 (4): 2343–94. doi:10.1021/cr3003533.

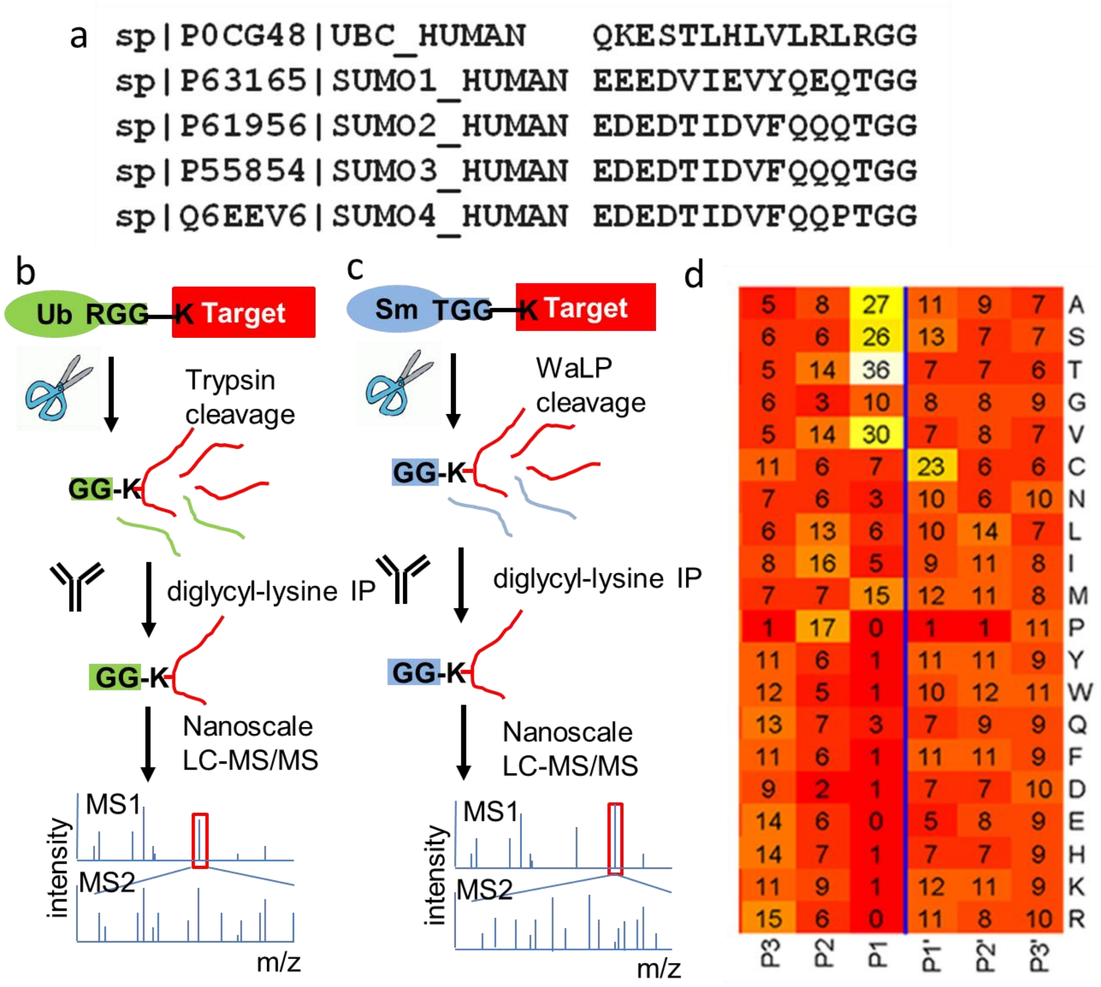
Chapter V

Global, site-specific identification and quantitation of endogenous human SUMO modifications

A. Introduction

The family of Small Ubiquitin-like Modifier (SUMO) proteins in humans includes four distinct genes with three types of members: SUMO1, SUMO2/3 (which differ by only three residues) and SUMO4. SUMO proteins regulate the function of various proteins by reversible, covalent, isopeptide-bond attachment between the C-terminus of SUMO and a free ε-amine group of the target protein's lysine (Makhnevych et al. 2009), similar to Ubiquitin (Ub). Ub conjugation mainly targets proteins for degradation by the proteasome, but has also been implicated in DNA repair, receptor signaling and cell communication (Heideker and Wertz 2015). The function of SUMO conjugation is less well-understood, but SUMOylated proteins are involved in gene expression, DNA repair, nuclear import, heat shock, cell motility, and lipid metabolism (Makhnevych et al. 2009; Flotho and Melchior 2013). SUMO targets are generally low-abundance proteins, and the amount of the modification at steady-state is also low (Becker et al. 2013). In recent years several groups have developed methods for analysis of SUMOylated proteins. However, many of these involve overexpression of mutant SUMO sequences. Therefore, methods that allow proteome-level identification of endogenous SUMOylation sites are needed.

Recently, a proteomic method has been developed to measure thousands of endogenous ubiquitination sites. The method takes advantage of the C-terminal sequence of Ub (RGG) (Figure 5.1). When cleaved with trypsin, peptides containing a Ub-remnant diglycyl-lysine are generated that can be enriched using specific



```
sp|P0CG48|UBC_HUMAN   QKESTLHLVLRLRGG
sp|P63165|SUMO1_HUMAN EEEDVIEVYQEQTGG
sp|P61956|SUMO2_HUMAN EDEDTIDVFQQQTGG
sp|P55854|SUMO3_HUMAN EDEDTIDVFQQQTGG
sp|Q6EEV6|SUMO4_HUMAN EDEDTIDVFQQPTGG
```

	P3	P2	P1	P1'	P2'	P3'
A	5	8	27	11	9	7
S	6	6	26	13	7	7
T	5	14	36	7	7	6
G	6	3	10	8	8	9
V	5	14	30	7	8	7
C	11	6	7	23	6	6
N	7	6	3	10	6	10
L	6	13	6	10	14	7
I	8	16	5	9	11	8
M	7	7	15	12	11	8
P	1	17	0	1	1	11
Y	11	6	1	11	11	9
W	12	5	1	10	12	11
Q	13	7	3	7	9	9
F	11	6	1	11	11	9
D	9	2	1	7	7	10
E	14	6	0	5	8	9
H	14	7	1	7	7	9
K	11	9	1	12	11	9
R	15	6	0	11	8	10

Figure 5.1. A strategy for mapping endogenous SUMO1/2/3 attachment sites. (a) C-terminal sequence alignment of mature human SUMO1-4 and Ubiquitin. (b) Cartoon showing ubiquitin-remnant mapping strategy. Proteins modified by ubiquitin are digested with trypsin leaving a diglycine attached to the epsilon-amine of the lysine where Ubiquitin was attached. An antibody specific for diglycine-modified lysine is used to enrich ubiquitin-remnant containing peptides that are then identified by nanoLC-MS/MS. (c) The same as (b) but using digestion with WaLP to generate the same diglycyl-lysine from SUMO attachment sites. (d) Heat map of the normalized percentage of each amino acid found at each position in peptides generated from a WaLP digest (Meyer et al. 2014).

antibodies and identified by tandem mass spectrometry (Figure 5.1b) (Xu et al. 2010; Kim et al. 2011; Wagner et al. 2011). Instead of RGG at the C-terminal, mature human SUMO isoforms have the sequence TGG and no tryptic cleavage site near the C-terminus (Figure 5.1a). Using various schemes introducing mutant SUMOs with tryptic cleavable sequences, several groups have developed methods for global profiling of SUMO attachment sites (Knuesel et al. 2005; Blomster et al. 2010). Last year, three groups reported global profiling approaches in which mutant SUMOs with various affinity tags and protease recognition sites were introduced into cells. Hendriks et al introduced a lysine deficient SUMO-2 with a C-terminal trypsin cleavage site, His10–SUMO-2 K0 Q87R, which is rendered resistant to Lys-C digestion. The SUMOylated proteins are enriched by immobilized-metal affinity chromatography (IMAC) and digested with Lys-C. Peptides modified with SUMO were then purified again with IMAC and finally digested with trypsin to generate a 5 amino acid C-terminal SUMO-remnant modification. This group used proteasome inhibition and heat shock to accumulate SUMOylated proteins, and were able to identify the sites of over 4,300 SUMO-acceptor lysines in over 1,600 proteins (Hendriks et al. 2014). Lamoliatte et al also introduced a mutant SUMO with C-terminal RN/EQTGG sequences and raised specific antibodies to each remnant. These researchers identified 954 sites on 538 proteins and were able to distinguish SUMO isoforms (Lamoliatte et al. 2014). Tammsalu et al introduced a His6-SUMO2 T90K and were able to identify 1002 unique SUMOylated sites in 539 proteins after heat shock (Tammsalu et al. 2014). Currently, no method exists for identifying endogenous SUMO sites on a global proteome scale without introduction of mutant SUMO.

We recently described the novel application of wild-type α-lytic protease (WaLP) to proteome digestion for shotgun proteomics. Although relatively relaxed specificity was observed, WaLP prefers to cleave after threonine residues and never cleaves after arginine (Meyer et al. 2014) (Figure 5.1d). We show here that WaLP cleaves at the C-terminal TGG sequence (in SUMO 2/3 and 4) leaving a SUMO-remnant diglycyl-lysine at the position of SUMO attachment. The resulting diglycyl-lysine-containing peptides can then be identified using methods already developed for Ub-profiling as described above (Fig 5.1c). The method allows identification of SUMO attachment sites under completely native conditions using the Ub-profiling workflow by simply substituting WaLP for trypsin. The same sample can be subjected to analysis of both Ub-attachment and SUMO-attachment simply by digesting the sample with either trypsin or WaLP respectively.

B. Materials and Methods

1. Cell culture and treatment

Human colorectal carcinoma HCT116 cells were grown in 245 mm X 245 mm plates using DMEM media supplemented with 10% FBS, penicillin, streptomycin and either 13C615N2 lysine (Cambridge Isotope Labs) or unlabeled lysine at 50 mg/L. Cells were treated with 10 μM MG132 (Sigma) dissolved in DMSO for 4 hours or DMSO-only negative control. Cells were then washed with PBS, lifted by scraping in ice-cold PBS, and counted with a TC20 cell counter (Biorad). Equal quantities of unlabeled and labeled cells from each condition were combined yielding 5 x 10^7 cells from each condition, which were washed with PBS on ice and then stored at -80 until lysis.

2. Cell lysis and digestion

Frozen cell pellets were thawed in 6 mls of lysis buffer containing 100 mM HEPES pH 7.2, 75 mM NaCl, 1% SDC, 1% SL, 1 mM Na3VO4, 1 mM β-glycerophosphate, 1 mM NaF, 2 mM NEM, 1 mM PMSF and Roche complete mini protease inhibitor. Cells were then sonicated on ice using 15w power output for 3 cycles of 30 seconds with 30 second rests in between. Insoluble material was precipitated by centrifugation at 20,000 x g for 15 minutes at room temperature and protein in the supernatant was quantified using the BCA assay. Each protein sample was then reduced with 4.5 mM DTT for 30 minutes at 55 ^{o}C. The reduction was quenched by addition of NEM to 10 mM for 30 minutes at room temp. A portion of each sample (20 mg of protein) was digested separately by addition of protease at 1:200, protease:proteome (either LysC + trypsin for ubiquitin profiling or WaLP for SUMO profiling). After digestion for 24 hours at 37 ^{o}C, the reaction was stopped by acidification with TFA to 1% and precipitated detergents were removed according to the method described previously (Lin et al. 2013). Peptide solutions were then filtered with 0.22 µm SFCA syringe filters and desalted using tC18 sepPak (500 mg, 6cc vac, Waters) eluted with 5 mls of 50% ACN in 0.5% acetic acid. A small aliquot from the SepPak elution (20 µg) was analyzed to ensure efficient digestion before IP. The eluted samples were frozen in liquid nitrogen and lyophilized.

3. Western blotting

Excess cells from the SILAC growths were lysed as described above for western blotting. Samples were boiled for 10 min after addition of 5X reducing SDS-PAGE buffer, separated on a 12% TGX gel (Biorad), and then transferred to PVDF using a Trans-blot semi-dry transfer apparatus (Biorad). Membranes were blocked for 1 hour at room temperature with 4% powdered milk in TBST, then incubated with primary antibodies overnight at 4 ^{o}C. HRP-conjugated secondary antibodies were incubated for 1 hr at room temp and detected using chemiluminescence (ECL, Amersham).

4. Off-line separation of peptides prior to IP

Samples were separated by basic-pH reversed phase (bRP) was performed using a 19mm ID X 25 cm long bridged-ethylene hybrid (BEH) C18 column with 5 µm particles (Waters). Mobile phases were 10mM ammonium formate pH 10, in water (phase A) or 10 mM ammonium formate in 90% ACN (phase B). Thirty-two fractions were collected and every fourth fraction was pooled into a final total of 4 fractions from each digest. Fractions were lyophilized to dryness, resuspended in 0.5% TFA, and desalted again with SepPak Vac tC18 cartridges (200 mg size). Desalted peptides were then lyophilized to dryness and stored at -80 until IP. diGlycyl-lysine Immunoprecipitation – Pre-coupled antibody resin (Ubi-scan from Cell Signaling Technologies) was cross linked using the Pierce crosslink IP kit according the manufacturer's protocol. For each biological condition, one tube of antibody-conjugated beads was split equally four aliquots (20 µl of beads per HPRP fraction), and was used per sample to immunoprecipitate the diGlycyl-lysine-containing peptides according to the CST protocol. Eluted peptides were desalted using RP30 desalting tips (Thermo-Fisher) and then analyzed by nLC-MS/MS on an orbitrap Fusion mass spectrometer at Harvard Medical School.

5. nLC-MS/MS

Mass spectrometry data were collected on an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) equipped with a Proxeon Easy nLC 1000 for online sample handling and peptide separations. Samples were resuspended in 8 µL of 5% formic acid + 5% acetonitrile and were loaded onto a 100 µm inner diameter fused-silica micro capillary with a needle tip pulled to an internal diameter less than 5 µm. The column was packed in-house to a length of 35 cm with a C18 reverse phase resin (GP118 resin 1.8 µm, 120 Å, Sepax Technologies). The peptides were separated using a 120 min linear gradient from 3% to 25% buffer B (100% ACN + 0.125% formic acid) equilibrated with buffer A (3% ACN + 0.125% formic acid) at a flow rate of 600 nL/min across the column. Precursor spectra were collected with a target resolution of 120,000 in the Orbitrap using a scan range of 300-2,000 m/z. The top 10 precursors with intensity greater than 5,000 were fragmented sequentially with either CID or ETD in the ion trap with the rapid scan rate, resulting in two separate spectra for each selected precursor ion.

6. Data Analysis

Peptides were first identified by database search with MS-GF+ trained for peptides from WaLP digestion. Two searches for each file were performed, one specifying fixed light lysine and one specifying fixed heavy lysine. All searches allowed variable oxidation of methionine, variable protein n-terminal methionine loss and acetylation at alanine or serine, variable peptide N-terminal pyro-glutamate from Q, variable diglycyl-lysine, and fixed modification of cysteine by N-ethyl maleimide. The .mzid output from fixed heavy lysine database searches were processed using R scripts (Team 2011) to combine the mass of heavy lysine and diglycine into one modification. All .mzid files were then converted to pepXML for compatibility with TPP (Keller et al. 2005) using idconvert.exe (Proteowizard, (Kessner et al. 2008)). PeptideProphet was used to refine heavy and light identifications separately (Keller et al. 2002). iProphet was used to combine files corresponding to HPRP fractions from a single condition, and to combine the results from separate heavy and light database searches (Shteynberg et al. 2011). PTMprophet was used to generate localization scores for diglycyl-lysine (Shteynberg et al. 2012; Shteynberg et al. 2013). Xpress was used to compute SILAC quantification (Han et al. 2001). The quantified Identifications were then filtered to <1% FDR, and the IDs were exported as an excel spreadsheet. Further processing was done with R. All identifications from the untreated/untreated distribution were used to calculate the standard deviation of the log Xpress ratio. The median of the log ratios was subtracted from each value to normalize the distribution. Peptides containing multiple SUMO-remnant modifications were excluded from quantitation. SUMO-remnant containing peptides were then filtered keeping only sites with localization scores >0.75. For protein modification sites identified by multiple peptides, the weighted-average of the Xpress ratio was computed using the total heavy and light area as weights. All unique SUMO modification sites were plotted and sites with log ratios outside 1 or 2 standard deviations were flagged as “changing”. The filtered peptide lists were output to excel.

All sites with log ratios over 1 standard deviation above the median were considered up-regulated, and all sites with log ratios more than 1 standard deviation below the median were considered down-regulated. Accessions were analyzed for GO-term enrichment and protein domain enrichment with DAVID (http://david.abcc.ncifcrf.gov/summary.jsp) using all human proteins as background.

C. RESULTS

1. Wild-type alpha-lytic protease (WaLP) digestion allows endogenous SUMO profiling

We reasoned that digestion of SUMOylated proteins with WaLP, which should cleave after the threonine in the SUMO C-terminal sequence, TGG, would generate a SUMO-remnant diglycyl-lysine at sites of SUMO attachment. Then the same workflow used for ubiquitin-remnant profiling could be used to globally profile SUMO attachment sites (Figure 5.1c). Although the WaLP can simply be substituted for trypsin during sample preparation, identification of non-tryptic peptides produced from WaLP digestion is challenging because search engines score on the basis of b and y ion series that are expected from tryptic peptides with a C-terminal positive charge. We showed that identification of peptides from WaLP digestion benefits from electron transfer dissociation (ETD) (Syka et al. 2004) , especially with precursor charge states > +2 (Meyer et al. 2014). Not only are the C-terminal residues generated from WaLP digestion not positively charged, WaLP cleaves after at least four different amino acids requiring the use of “no enzyme” specificity in database searches. We found that MS-GF+ (Kim and Pevzner 2014) outperforms many other search engines (Meyer et al. 2014). For the SILAC data presented here, MaxQuant (Cox et al. 2011) resulted in very few peptide identifications as compared to MS-GF+. MS-GF+ is also very fast and the scoring function can be trained using identifications from an initial search (Kim and Pevzner 2014).

2. Global SUMO profile results

Carr's group optimized the protocol for purifying and immunoprecipitating (IP) diglycyl-lysine-containing peptides and reported that pre-IP fractionation of peptides from tryptic digests with basic pH reversed phase (bRP) chromatography improved identification of ubiquitin-remnant peptides (Udeshi et al. 2013). We followed the protocol from the Carr group exactly, and also found that pre-IP bRP fractionation of peptides from WaLP digestion more than doubled the number of identified SUMOylation sites as compared with the standard IP protocol. We refer to these samples as the bRP-IP sample and the IP sample respectively. The samples were analyzed on a Thermo Fusion mass spectrometer. The top 10 precursors with intensity greater than 5,000 were fragmented sequentially with either CID or ETD in the ion trap with the rapid scan rate, resulting in two separate spectra for each selected precursor ion. We identified 707 unique SUMO attachment sites in 443 proteins (Table 5.1) from a SILAC experiment (Ong et al. 2002) comparing a MG132-treated and untreated cells. The entire sample set resulted in 22,933 unique peptides identified, of which 2, 412 were unique diglycyl-lysine-containing peptides. Of these

Table 5.1. Summary of identified and quantified peptides from both the bRP-IP experiment and from the standard IP experiment.

	Unique peptides	**Digly PSMs**	**Digly unique peptides**	**unique digly sites localized**
bRP identified	22,933	2,412	1,840	707
bRP quantified	14,356	1,849	1,407	589
standard IP identified	5,997	1,055	897	318
Standard IP quantified	3,763	787	670	247

2,412 SUMO-remnant peptide spectrum matches, 741 resulted from CID, and 1,671 (69%) resulted from ETD consistent with our previous findings on the effectiveness of ETD for sequencing the non-tryptic peptides resulting from WaLP digestion (Meyer et al. 2014). Of the 2,412 peptide-spectrum matches, 1,840 corresponded to unique diglycyl-lysine-containing peptides which were confidently localized (PTMprophet score >0.75) to 707 unique sites.

Since WaLP cleaves at different residues than trypsin, we examined the overlap of our site identifications with those recently reported by Hendricks' et al. (Hendriks et al. 2014). Only 39% of our 707 identified SUMOylation sites overlap with the 4,910 sites amalgamated by Hendricks et al. (Hendriks et al. 2014). The sites were also compared to SUMO sites reported in Phosphosite Plus (Figure 5.2a) (Hornbeck et al. 2012). When compared to all previously reported SUMOylation sites including those in Phosphosite Plus, we found 293 of the 707 sites we found were previously reported, and 414 have not been previously reported. Peptides containing SUMO-remnants from WaLP digestion may correspond to sequences that cannot be covered by tryptic digestion due to the abundance or lack of nearby tryptic cleavage sites (Meyer et al. 2014). We found at least two examples of such cases (Figure 5.2b, 5.2c). One corresponded to a SUMOylation site would reside on a tryptic peptide of length 30 (Figure 5.2b), which is unlikely to be identified with traditional bottom-up proteomic methods. Another (Figure 5.2c) is in an arginine-rich region of SNUT2 that would be cleaved into a peptide of length 2. Multiple SUMOylation sites were found in 60% of the proteins identified, with 17% having two, 7% having 3, and 4.6% having five or more.

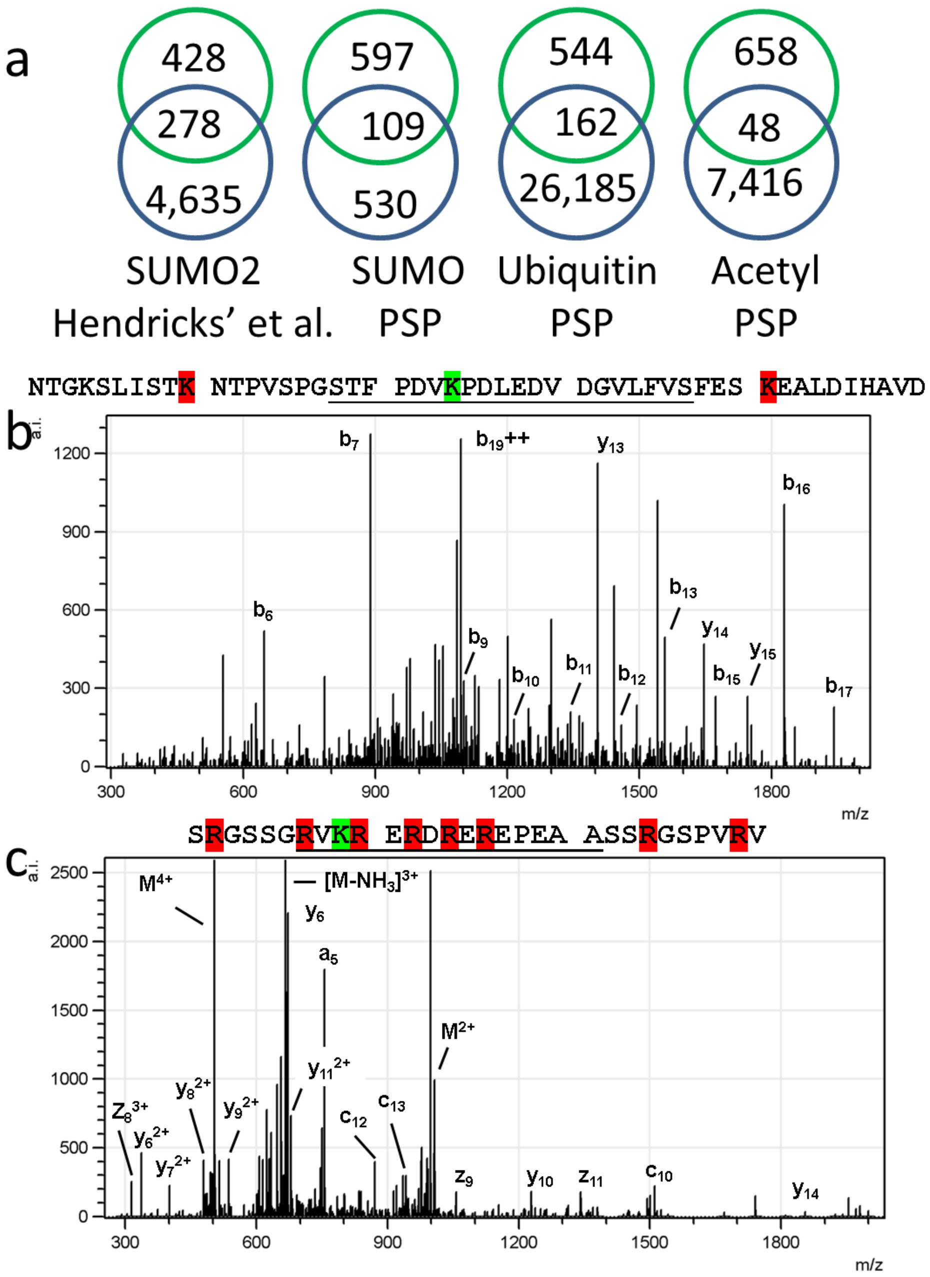


Figure 5.2. Summary of identifications from this study and comparison with previous studies. (a) Venn diagrams showing overlap between all SUMO sites identified in this study and all known lysine modifications from either Hendricks et al(Hendriks et al. 2014) or Phosphosite Plus (Hornbeck et al. 2012). (b) Example spectrum of a novel site identified from MGAP that is unlikely to be found after digestion with LysC or trypsin. The sequence above the spectrum shows the tryptic cleavage sites in red, the modification site in green, and the identified sequence underlined. (c) Identified ETD spectrum of a SUMO-remnant peptide corresponding to position 29 of SNUT2. The sequence above the spectrum shows this position is surrounded by tryptic cleavage sites.

To look for sites involved in PTM crosstalk, we compared the sites we identified with lysine methylation, acetylation, and ubiquitylation sites reported in Phosphosite Plus (Hornbeck et al. 2012). No overlap was found between our identified sites and methylation sites. Of our identified SUMOylation sites, 7% matched known acetylation sites, and 29% of our identified sites overlap with previously reported ubiquitination sites. These results are very similar to those previously reported by Hendricks et al., who found 8% of their SUMOylation site IDs overlap with acetylation sites and 22% of their identified SUMO sites overlap with known ubiquitination sites (Hendriks et al. 2014).

3. SILAC quantitation of SUMO site changes in MG132 treated cells

From the bRP-IP sample, 589 sites out of the 707 unique sites identified could be quantified, and from the IP sample, 247 sites out of 318 sites could be quantified (of which 191 overlapped with the 589 from the bRP-IP sample). The results were normally distributed, and 37% of the quantification values fell outside at least 1 standard deviation of the median (Figure 5.3a). Nearly equal numbers of sites were significantly upregulated (109) as were down-regulated (111) by MG132 treatment similar to previous reports (Hendriks et al. 2014). The large numbers of sites that changed in abundance is not likely due to changes in protein levels as protein levels were not found to change significantly during a 4 hr MG132 treatment (Kim et al. 2011). Pearson correlation analysis of the abundance of peptides corresponding to the 191 sites quantified in biological replicates enriched with either the standard IP or pre-IP fractionated gave a slope of 0.93 +/- 0.05, and an R of 0.81, also similar to what was reported previously (Hendriks et al. 2014) (Figure 5.3b).

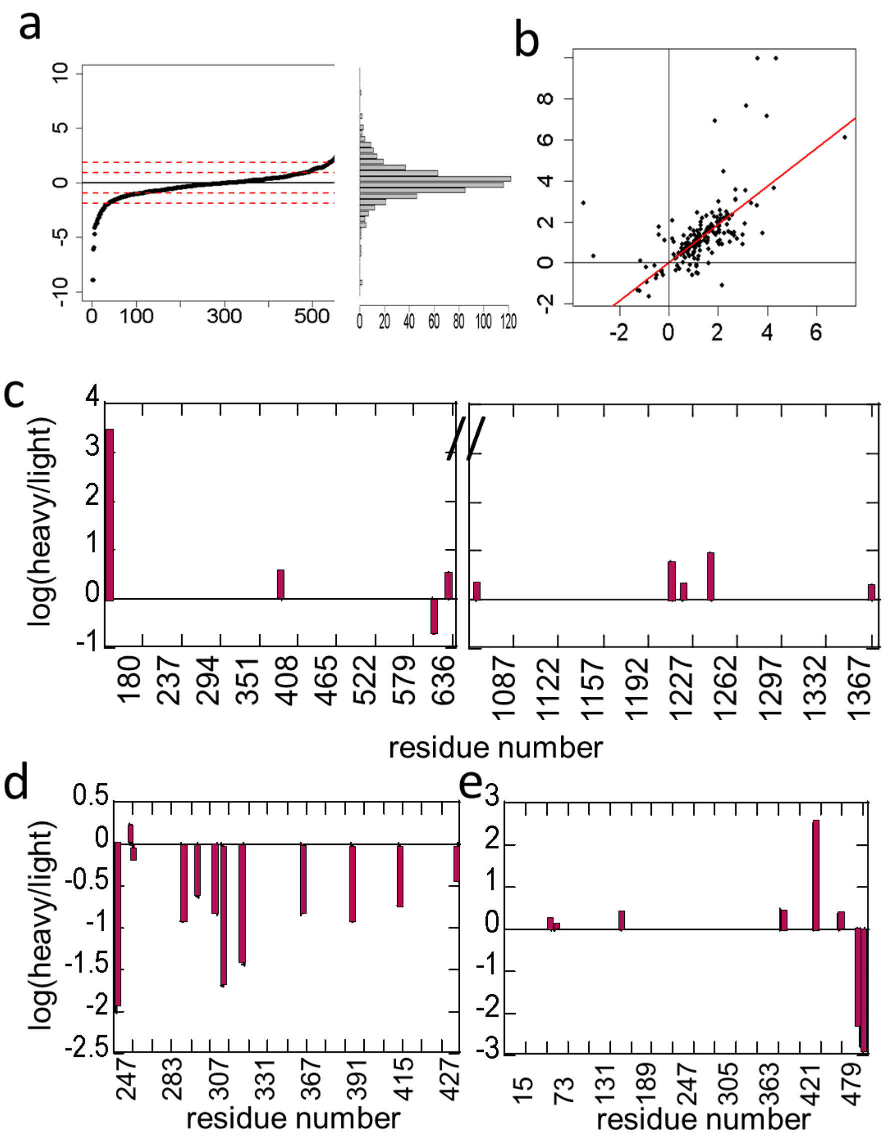


Figure 5.3. Results for all unique SUMO sites quantified by SILAC in this study. (a) Plot of normalized log2 ratios (heavy/light) for all unique sites quantified from the bRP prefractionated sample plotted against their arbitrary index number (b) Correlation of measured abundance (log2 ratios (heavy/light) integrated peak areas) of diglycyl-lysine containing peptides quantified in both the pre-IP bRP fractionated sample and compared to the standard IP sample. The slope of 0.93 +/- 0.05 and the R value of 0.81 indicate the reproducibility of the results. (c) Change in abundance of the 12 SUMOylation sites detected in remodeling and spacing factor (uniprot Q96T23). (d) Only one of the nine SUMOylation sites detected in topoisomerase 2A (uniprot P11388) was significantly changed upon MG132 treatment. (e) Of the nine SUMOylation sites detected in PML (uniprot P29590), six were unchanged, one was significantly increased and two were significantly decreased upon MG132 treatment.

Interestingly, for those proteins containing multiple sites, all sites were not regulated similarly upon proteasome inhibition. None of the 12 sites in remodeling and spacing factor (uniprot Q96T23) were significantly changed. Topoisomerase 2-alpha contained the greatest number of quantified SUMOylation sites, 11/12 of which decreased in SUMOylation levels upon treatment with MG132 (Figure 5.3c). In contrast, 9 sites quantified on remodeling and splicing factor 1 nearly all increased upon MG132 treatment (Figure 5.3d). Of the nine SUMOylation sites detected in PML (uniprot P29590), six were unchanged, one was significantly increased and two were significantly decreased upon MG132 treatment (Figure 5.3e). An interesting example was the thyroid hormone receptor (uniprot Q9Y2W1) in which four SUMOylation sites were identified. The SUMOylation at position 592 significantly increased whereas SUMOylation at position 603 significantly decreased. Similarly the SUMOylation at position 697 significantly increased whereas SUMOylation at position 705 significantly degreased. These results suggest that, similar to ubiquitylation events (Kim et al. 2011), not all SUMOylation events on target proteins are regulated identically.

We next extracted sequence windows surrounding the site of sumo modification to examine trends in SUMO attachment motifs (Fig. 5.4). As reported consistently, we found that 43 % of the SUMO attachments occurred at the sequence motif (ΨKXE/D), 15% corresponded to the inverted consensus (E/DXK), and 42% did not correspond to either consensus (Figure 5.4a-b). We further examined the subsets of sites that were quantitatively up or down for trends in attachment upon proteasome

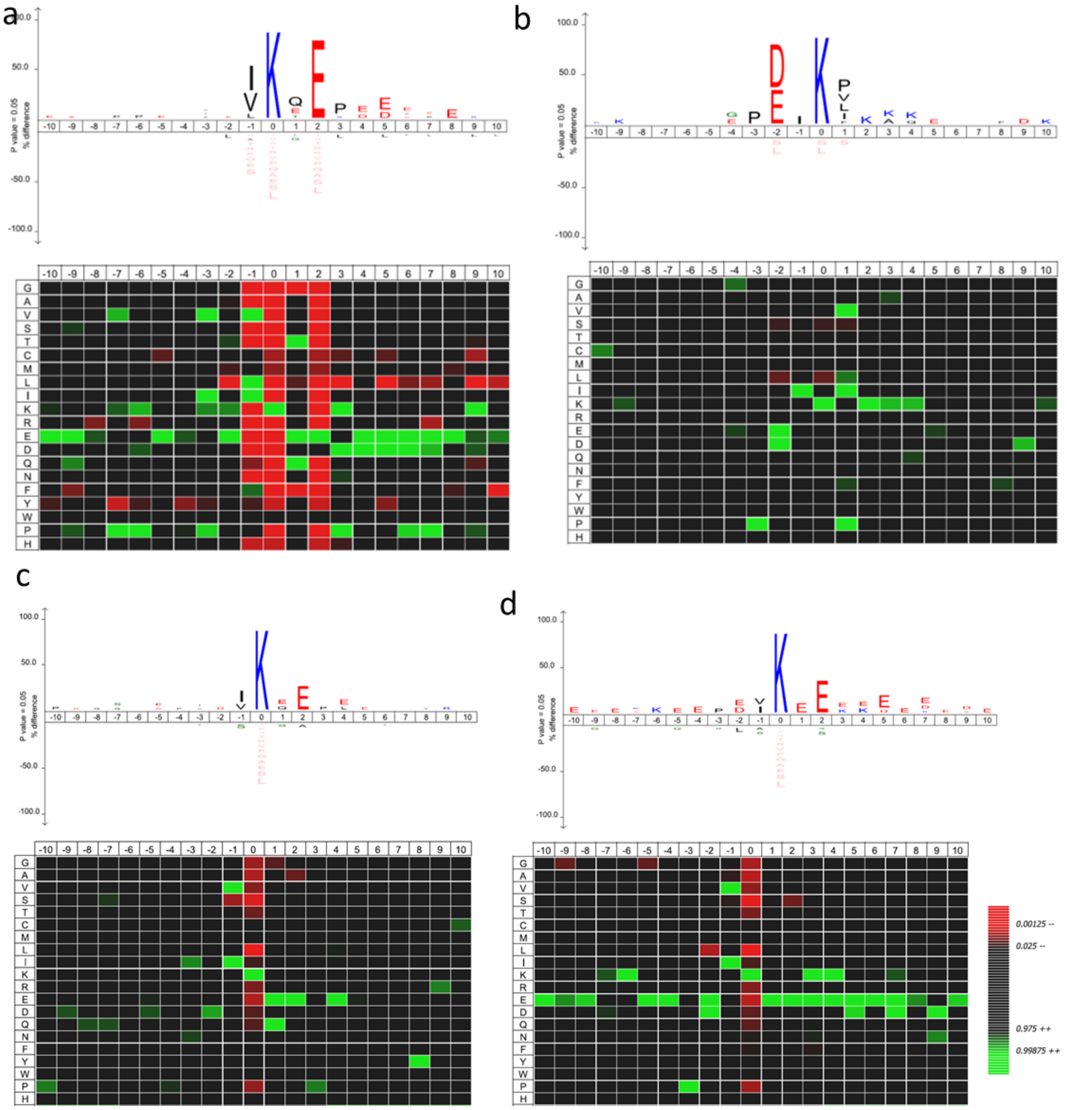


Figure 5.4. Motif analysis of all unique SUMO sites identified in this study. (a) SubLogos and heatmaps of sites separated by the forward motif (KXD/E), (b) or the inverted motif, or neither forward/inverted. (c) Motif analysis of only sites found to increase upon MG132 treatment. (d) Motif analysis of only sites found to decrease upon MG132 treatment.

inhibition by MG132 (Figure 5.4c, d). Such trends may be indicative of E3 activity upon proteasome inhibition. Out of the 109 motifs that had increased SUMOylation upon MG132 treatment, 37% corresponded to the forward motif, 19% to the reverse motif, and 44% to neither motif. Thus, the up-regulated motifs reflected the total distribution. The motif analysis of the 111 SUMO attachment sites that were down-regulated by MG132 treatment showed 42% correspond to the forward motif, 30% to the reverse motif, and 28% to neither motif. Thus, those sites that were significantly down regulated were enriched in the reverse motif. In addition, the sequence surrounding the down-regulated motifs appears to be enriched in acidic residues (Figure 5.4d).

Gene ontology term enrichment was performed to understand biological significance of the proteins that increase or decrease after proteasome inhibition (Figure 5.5). We found four known SUMO E3 ligases, PIAS1, PIAS3, RBP2, and TOPRS, which all contained SUMOylation sites that increased in occupancy upon treatment with MG132.

4. Validation of biological results

We validated four of the novel proteins found to be SUMOylated by western blot, SPT3, SFPQ, Syne-1, and WSTF (Figure 5.6). SUMO1/2/3 was IPed and the novel target proteins were detected by western blot.

5. Discussion

We developed a method for global profiling of SUMOylation events that would allow detection of native SUMO taking advantage of the specificity of WaLP

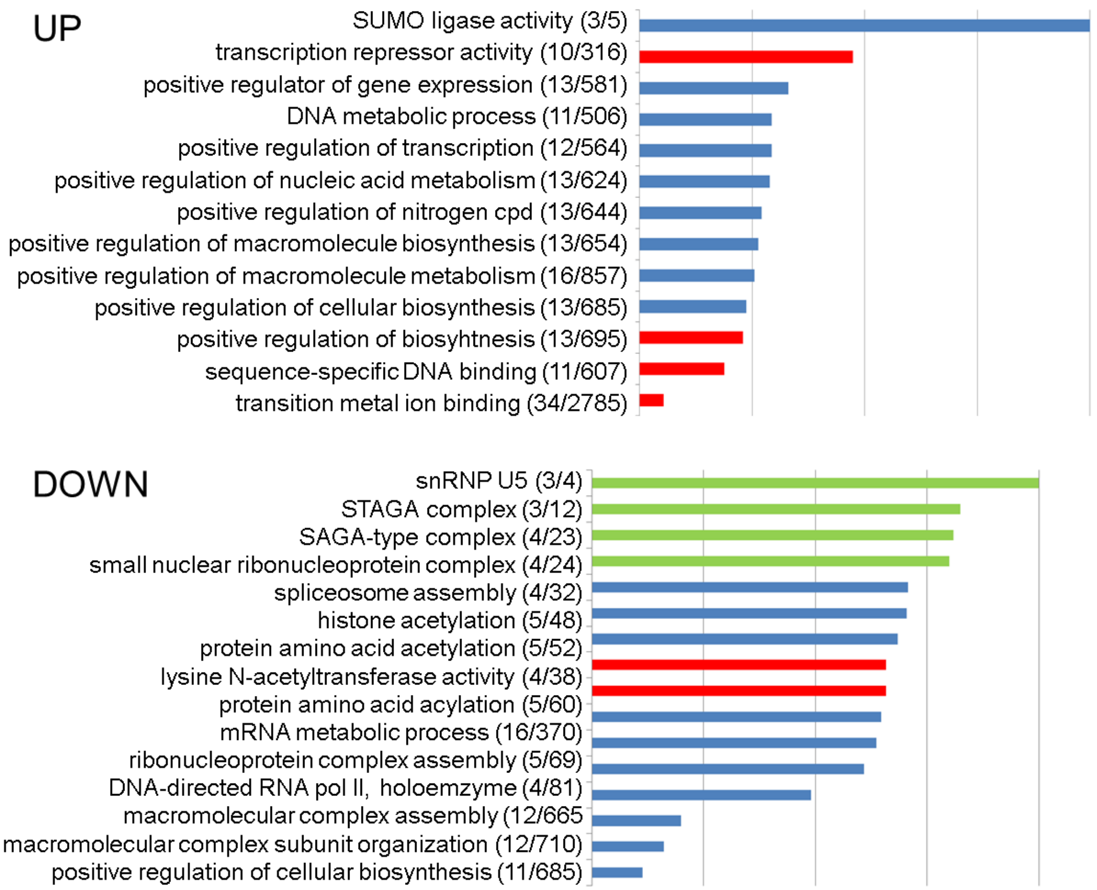


Figure 5.5. GO term enrichment analysis. Analysis of all the proteins in which SUMO sites were quantified revealed some GO categories segregated according to whether the SUMOylation site was up- or down-regulated. That is, all of the SUMOylation sites in proteins that fall into these categories are regulated in the same way. Although many GO categories did not segregate, those that did segregate are plotted. The up-regulated categories includeBlue bars correspond to GO biological processes, red bars correspond to molecular functions and green bars correspond to cellular compartments.

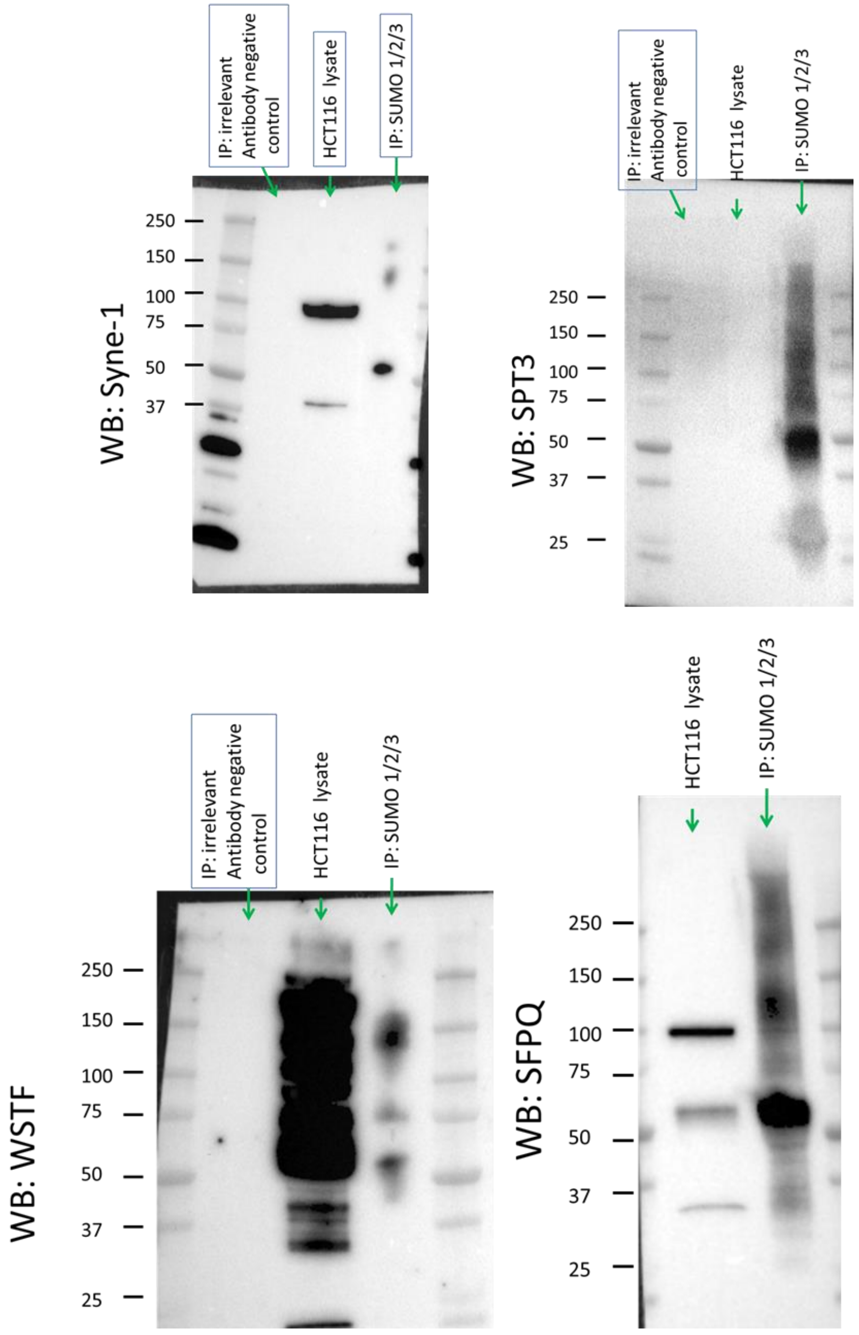


Figure 5.6. Validation of novel SUMOylated proteins by western blot. SUMO1/2/3 was immunoprecipitated and the resulting proteins were western blotted for the SUMOylated target protein alongside the whole cell lysate and a negative control IP with an irrelevant antibody.

for cleavage after threonine. By simply substituting WaLP for trypsin it was possible to immunopurify and identify a large number of diglycyl-lysine containing peptides corresponding to SUMO remnants. Surprisingly, a large number of the identified sites corresponded to novel SUMOylation sites. Several reasons could explain the large number of new sites identified. First, as we reported previously, the orthogonal specificity of WaLP allows cleavage of proteins at sites that may not be accessible to trypsin (Meyer et al. 2014). Second, although previous studies attempted to achieve minimal expression of their mutant SUMO construct, it is possible that slight overexpression of SUMO or the presence of mutant sequences could cause unnatural SUMO attachment. Third, our method does not differentiate between SUMO isoforms 1-4, whereas Hendricks et al. examined only SUMO-2 attachment (Hendriks et al. 2014).

Confidence regarding the new site identifications was garnered from IceLOGO motif calculations which gave identical distributions of forward and reverse sequence motifs as have been reported previously (Hendriks et al. 2014). In addition, the proteins to which we found SUMOs attached corresponded to the expected GO terms such as transcription regulation and RNA processing (Lamoliatte et al. 2014).

To understand whether SUMOylation of certain sites changed upon MG132 treatment, we calculated the IceLOGO motifs for only the sites that were significantly increased or decreased upon MG132 treatment. The sites that significantly increased showed no differences in the distribution of forward, reverse and other motifs. Interestingly, the sites that were significantly decreased were more strongly represented by the reverse motif increasing from 19% to 30% of the motifs. We also analyzed the GO terms associated with only those proteins in which sites increased or decreased significantly. Although a number of GO terms were represented by sites that increased, decreased or stayed the same, some GO terms segregated completely. That is, all of the sites in proteins associated with that GO term appeared to be regulated similarly. Proteins functioning in transcriptional regulation, DNA synthesis, DNA binding and macromolecular biosynthesis segregated with increased SUMOylation upon MG132 treatment. GO terms associated with decreased SUMOylation were related to RNA splicing and acetyltransferase activity.

Another powerful aspect of the method is that it allows simultaneous determination of ubiquitylation and SUMOylation in the same sample. The same population of cells or tissue can be subjected to analysis of both Ub-attachment and SUMO-attachment simply by splitting the sample in two and digesting half with trypsin and the other half with WaLP. The samples can then be processed in parallel to purify the peptides by bRP chromatography, di-gly IP, and mass spectrometric analysis. For the mass spectrometry, it is best to use optimized ionization approaches and data analysis tailored to the non-tryptic WaLP peptides. Finally, since the commercial antibody used in this study has been previously used for enrichment of Ubiquitin-remnant peptides, there may be value in developing additional antibodies against diglycine-remnants on known SUMOylation motifs.

D. References

Becker, J., S. V. Barysch, S. Karaca, C. Dittner, H. H. Hsiao, M. Berriel Diaz, S. Herzig and M. F. Urlaub H (2013). "Detecting endogenous SUMO targets in mammalian cells and tissues." Nat Struct Mol Biol: 525-531.

Blomster, H. A., S. Y. Imanishi, J. Siimes, J. Kastu, N. A. Morrice, J. E. Eriksson and L. Sistonen (2010). "In vivo identification of sumoylation sites by a signature tag and cysteine-targeted affinity purification." J Biol Chem **285**: 19324–19329.

Cox, J., N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen and M. Mann (2011). "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment." J Proteome Res **10**(4): v.

Flotho, A. and F. Melchior (2013). "Sumoylation: A Regulatory Protein Modification in Health and Disease." Annu Rev Biochem **82**: 357-385.

Guthals, A., C. Boucher and N. Bandeira (2014). "The Generating Function Approach for Peptide Identification in Spectral Networks." J Comput Biol **in press**.

Han, D. K., J. K. Eng, H. Zhou and R. Aebersold (2001). "Quantitative profiling of differentiation-induced microsomal proteins using isotope- coded affinity tags and mass spectrometry." Nature Biotech **19**: 946-951.

Heideker, J. and I. E. Wertz (2015). "DUBs, the regulation of cell identity and disease." Biochem J **465**(1): 1-26.

Hendriks, I. A., R. C. D'Souza, B. Yang, M. Verlaan-de Vries, M. Mann and A. C. Vertegaal (2014). "Uncovering global SUMOylation signaling networks in a site-specific manner." Nat Struct Mol Biol **21**(10): 927-936.

Hornbeck, P. V., J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham and M. Sullivan (2012). "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse." Nucleic Acids Res **40(Database issue)**: D261-270.

Keller, A., J. Eng, N. Zhang, X. Li and R. Aebersold (2005). "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." Mol Syst Biol **1**: 2005.0017.

Keller, A., A. I. Nesvizhskii, E. Kolker and R. Aebersold (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Anal Chem **74**: 5383–5392.

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

Kim, S. and P. A. Pevzner (2014). "MS-GF+ makes progress towards a universal database search tool for proteomics." Nat Commun **5**(5277): 1-10.

Kim, W., E. J. Bennett, E. L. Huttlin, A. Guo, J. Li, A. Possemato, M. E. Sowa, R. Rad, J. Rush, M. J. Comb, J. W. Harper and S. P. Gygi (2011). "Systematic and quantitative assessment of the ubiquitin-modified proteome." Mol Cell **44**(2): 325-340.

Knuesel, M., H. T. Cheung, M. Hamady, K. K. Barthel and X. Liu (2005). "A method of mapping protein sumoylation sites by mass spectrometry using a modified small ubiquitin-like modifier 1 (SUMO-1) and a computational program." Mol Cell Proteomics **4**: 1626–1636.

Lamoliatte, F., D. Caron, C. Durette, L. Mahrouche, M. A. Maroui, O. Caron-Lizotte, E. Bonneil, M. K. Chelbi-Alix and P. Thibault (2014). "Large-scale analysis of lysine SUMOylation by SUMO remnant immunoaffinity profiling." Nat Commun **5**.

Lin, Y., L. Huo, Z. Liu, J. Li, Y. Liu, Q. He, X. Wang and S. Liang (2013). "Sodium Laurate, a Novel Protease- and Mass Spectrometry-Compatible Detergent for Mass Spectrometry-Based Membrane Proteomics." PLoS ONE **8**(3): e59779.

Makhnevych, T., Y. Sydorskyy, X. Xin, T. Srikumar, F. J. Vizeacoumar, S. M. Jeram, Z. Li, S. Bahr, B. J. Andrews, C. Boone and B. Raught (2009). "Global map of SUMO function revealed by protein-protein interaction and genetic networks." Mol. Cell **33**: 124–135.

Meyer, J. G., S. Kim, D. A. Maltby, M. Ghassemian, N. Bandeira and E. A. Komives (2014). "Expanding proteome coverage with orthogonal-specificity α-lytic proteases." Mol Cell Proteomics **13**(3): 823-835.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-386.

Shteynberg, D., E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold and A. I. Nesvizhskii (2011). " iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates." Mol Cell Proteomics **10**: M111.007690.

Shteynberg, D., A. I. Nesvizhskii, R. L. Moritz and E. W. Deutsch (2013). "Combining results of multiple search engines in proteomics." Mol Cell Proteomics **12**(9): 2383-2293. .

Shteynberg, D. D. E., L. Mendoza, J. Slagel, H. Lam, A. Nesvizhskii and R. Moritz (2012). "PTMProphet: TPP Software for Validation of Modified Site Locations on Post-Translationally Modified Peptides." 60th American Society for Mass Spectrometry (ASMS) Annual Conference, Vancouver, Canada.

Syka, J. E. P., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." Proc Nat Acad Sci USA **101**(26): 9528 -9533.

Tammsalu, T., I. Matic, E. G. Jaffray, A. F. M. Ibrahim, M. H. Tatham and R. T. Hay (2014). "Proteome-Wide Identification of SUMO2 Modification Sites." Science Signaling **7**(323): rs2 1-10.

Team, R. D. C. (2011). "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0: URL http://www.R-project.org/.

Udeshi, N. D., P. Mertins, T. Svinkina and S. A. Carr (2013). "Large-scale identification of ubiquitination sites by mass spectrometry." Nat Protoc **8**(10): 1950-1960.

Wagner, S. A., P. Beli, B. T. Weinert, M. L. Nielsen, J. Cox, M. Mann and C. Choudhary (2011). "A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles." Mol Cell Proteomics **10**: M111 013284.

Xu, G., J. S. Paige and S. R. Jaffrey (2010). "Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling." Nat Biotechnol **28**: 868–873.

Appendix A

Analysis of MS data from SUMO-remnant IP

A. Database searches

Instrument .RAW files are first converted to .mzXML using Trans-proteomic Pipline (TPP, downloaded from http://sourceforge.net/projects/sashimi/files/latest/download?source=files). MS-GF+ is then used for database searching (downloadable from:http://proteomics.ucsd.edu/software-tools/ms-gf/). All files mentioned without explicit locations can be obtained from: https://github.com/jgmeyerucsd/SUMO-remnant.

1. Convert .RAW files to mzXML files using TPP.

First, you need to move your data files to: c:/Inetpub/wwwroot/ISB/data/(subdirectories are ok in here). Access TPP from the shortcut, login as guest/guest

From the HOME tab, click on "Analysis Pipeline,"and choose "Thermo RAW." Click on "ADD FILES"and then navigate using the buttons on the right to your data, click the boxes corresponding to your data files, and choose select at the bottom. Under section 3 (conversion options), choose "convert to mzXML," and then click on the button that says convert to mzML. The program will write the converted files into the same folder as your RAW files.

2. Database searching data from WaLP digestion using MSGF+

MSGF+ is run in batch mode once the program is installed on your computer. There is a directory on the desktop of .bat files (and these are also available on the github). You will need to edit the appropriate .bat file for your database search.An example batch file has been made available for database searching Orbitrap Fusion data collected on peptides generated from WaLP digestion. This file is titled, "MSGFplus_example_WaLP_search.bat," and this file will perform database searches on the four files from the bRP-IP experiment of MG132-treated HCT116 cells. Open the .bat file in notepad, and edit the file names and locations. Start the database search by saving the batch file and double clicking on it.

All database search parameters are contained in the .bat file, so it is worthwhile looking through it and understanding what each command does. Run MSGFplus.jar without commands to print descriptions of each command. Note that the correct training data is called by the –e command which specifies the enzyme (trypsin or WaLP and there are several different options for training data sets for WaLP) as is listed in the enzymes.txt file in the params folder (located in the F:/msgfplus.20140716/). When –e is specified as 10, this will call the newer training data for the Fusion analysis of WaLP peptides.

3. Database Searching Tryptic Data

An example windows batch file for MS-GF+ database searching of data from trypsin digestion is titled "exampleMSGFplus_search.bat" that contains the following text:

```
:: change directory
F:
cd \msgfplus.20140716\
:: the line below just gives the options, press any key to start the search
java -Xmx12000M -d64 -jar MSGFplus.jar
```

```
pause
:: each option in the lines below must be set according to the options printed from the
above command depending on the specific type of data
java -Xmx12000M -d64 -jar MSGFplus.jar -thread 8 -s
C:\Inetpub\wwwroot\ISB\data\exactive\2015\mg132\sbl_009779.mzXML -o
output\exactive\2015\Mg132\mg132pl_4_light1.mzid -e 1 -ti 0,2 -tda 1 -inst 1 -ntt 0 -
m 0 -t 10ppm -d database\110712_human.cc.fasta -mod digly.txt
java -Xmx12000M -d64 -jar MSGFplus.jar -thread 8 -s
C:\Inetpub\wwwroot\ISB\data\exactive\2015\mg132\sbl_009779.mzXML -o
output\exactive\2015\Mg132\mg132pl_4_heavy1.mzid -e 1 -ti 0,2 -tda 1 -inst 1 -ntt 0 -
m 0 -t 10ppm -d database\110712_human.cc.fasta -mod diglyh.txt
pause
```

The above batch file will run two database searches of one file from the tryptic data and output the .mzid file at the location specified after the tag “-o”.

4. Fix the formatting in your output files from the heavy database search

Database search output files from the heavy database search in the format of .mzID must next have their modification format updated. First you need to open the mzIDfix.R file in the R editor and change the file name and locationin the setwd line. You find this file by double clicking on the folder icon, then nagivate to Documents/GitHub/SUMO_remnant and then double click on the mzIDfix.R file. Then you have to set the file number to the number that the file is in the list, so if you have three files, enter the name of the first one and the number 1, for the second file, you should enter the name of the second file and the number 2, etc. Also change the output name at the bottom of the script. You should make sure only .mzid files are in the directory. To run the script, highlight all of it and right click and hit run script. You need to run this script for each file individually, each one takes about 20 minutes to run, so be patient.

Here is the code for the mzIDfix.R:

```
### the below line should be the directory where your mzid file is
setwd("C:/Users/JgMeyer/Documents/R/pepsum/pepsum")
### make an R object containing file names that have mzid suffix
files<-list.files(pattern="mzid")
#### print the files in the above objectfiles
### determine which file number from those printed is the file you want to fix
### load required external libraries
library(mzID)
library(XML)
### parse and find peptide elements
### set the number in brackets to the correct file number in the "files" object
doc <- xmlParse(files[3])
r = xmlRoot(doc) #gives content of root
### get the peptide tags and store in tags object
tags <- xmlElementsByTagName(r[[3]], "Peptide")
#### function to fix all tags, run all lines in order
```

```
manipulate <- function(tag) {
   ## get 'Modification' node set
   dMod <- tag["Modification"]
   ## get 'location' numbers
   loc <- sapply(dMod, xmlGetAttr, "location")
   ## get the sum of 'monoisotopicMassDelta'
   lapply(unique(loc), function(i) {
      if(length(d <- dMod[loc == i]) > 1) {
         nm <- "monoisotopicMassDelta"
         s <- sapply(d, xmlGetAttr, nm)
         xmlAttrs(d[[1]])[nm] <- sum(as.numeric(s))
      }
   })
   ## remove duplicated location nodes
   removeNodes(dMod[duplicated(loc)])
   ## return the adjusted tag
   tag
}
#### this command takes about 20 minutes, R appears to freeze but it does not
###actually, just let it work until you can interact with the GUI again
lapply(tags, manipulate)   ### applies manipulate to the XML structure to fix tags
#### saves new XML file, change the text equal to file to change the file output name
saveXML(r, file="dh_fix.mzid", indent=TRUE)
```

5. Convert each mzID file into pep.xml

Each mzID file can then be converted to pep.xml using idconvert.exe (version is 3.0.4388, newer versions have issues).This program is located in C:\idconvert\. Before running the idconvert.exe file on one of the fixed files, you need to paste the xml header into the first line. There is a word document called XMLHeader.txt in the SUMOanalysis folder on the desktop. For ALL the WaLP data files, you also need to replace the enzyme info with trypsin since WaLP is not recognized by the idconvert as an enzyme. There is text in the SUMOanalysis folder file titled idconvert_enzymetext.txt that shows you what the offending text is to search for and what to replace it with.

Find

```
<Enzymes>
<Enzyme missedCleavages="1000" semiSpecific="true" id="WaLP">
<EnzymeName>
<userParam name="wt-aLP"/>
```

Replace with

```
<Enzyme missedCleavages="1000" semiSpecific="true" id="Tryp">
<EnzymeName>
<cvParam accession="MS:1001251" cvRef="PSI-MS" name="Trypsin"/>
```

NOW you can run idconvert successfully. For example, in the cmd window, type

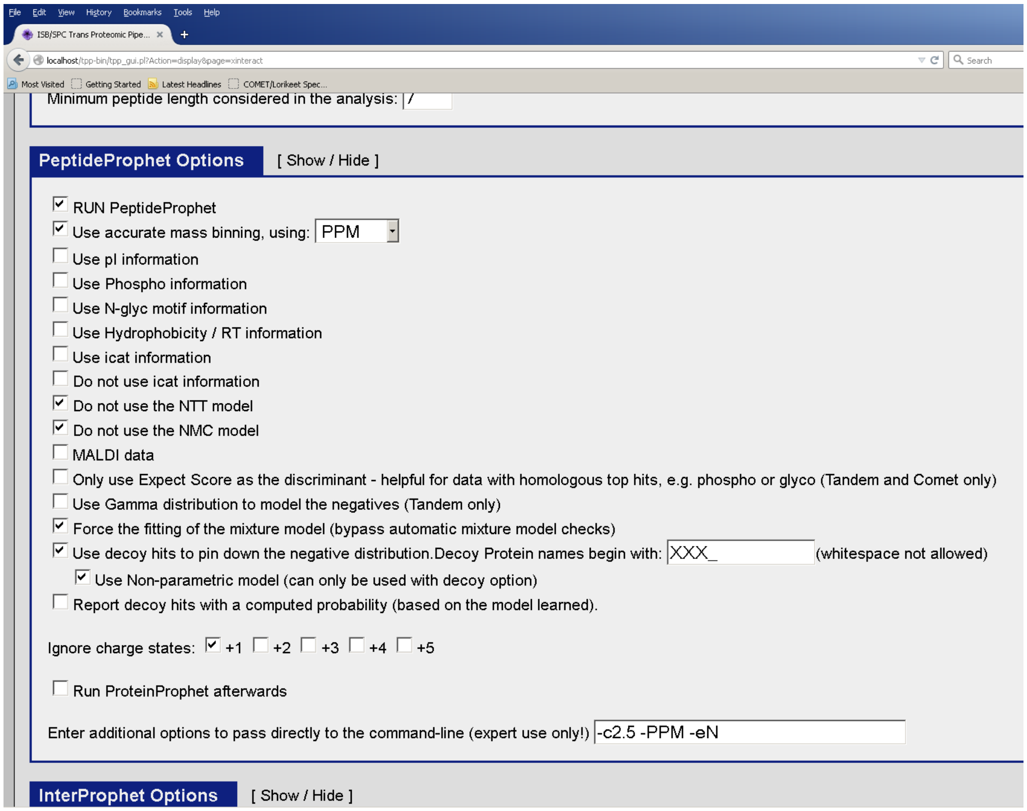
```
cd C:\idconvert
idconvert.exe F:\MSGFPlus.20140716\output\test\wh_fix.mzid –v –e .pep.xml –o
F:\\MSGFPlus.20140716\output\test\
```

This command will convert the wh_fix.mzid file into a wh_fix.pep.xml file and place it in the F:\\MSGFPlus.20140716\output\test\ directory.You will find this text also in the SUMOanalysis folder so you can edit it for your file name. It's called idconvert_cmdlines.txt. Once you've edited it, copy, right click and paste into the cmd window.

B. TPP refinement of identifications

1. Run Peptide Prophet

The .pep.xml files are then refined for each database search output file individually with PeptideProphet within the TPP graphical user interface (TPP-GUI, open the program from the start menu, see also the extensive TPP documentation and walkthrough here: http://www.proteomecenter.org/software.php). Move your files (the .pep.xml and the .mzXML files)to any folder within the directory C:\Inetpub\wwwroot\ISB\data\, or create your own subfolder. Login as guest, mouse-over "analysis pipeline," and click on "analyze peptides." Choose your .pep.xml files using the button that says "add files."You will want to have all the files that you eventually plan to combine in the same folder, and analyze them at the same time. Check the boxes to the right in the "Select Files to Analyze" window. Be sure to check the box that says, "process each file individually." Leave "write output to file"as the default (interact.pep.xml), and it will name your output files as their original name plus interact. The parameters used for PeptideProphet are shown in Figure 6.1.


File Edit View History Bookmarks Tools Help
ISB/SPC Trans Proteomic Pipe...
localhost/tpp-bin/tpp_gui.pl?Action=display&page=xinteract
Search
Most Visited
Getting Started
Latest Headlines
COMET/Lorikeet Spec...

Minimum peptide length considered in the analysis: 7

PeptideProphet Options [Show / Hide]

☑ RUN PeptideProphet
☑ Use accurate mass binning, using: PPM
☐ Use pI information
☐ Use Phospho information
☐ Use N-glyc motif information
☐ Use Hydrophobicity / RT information
☐ Use icat information
☐ Do not use icat information
☑ Do not use the NTT model
☑ Do not use the NMC model
☐ MALDI data
☐ Only use Expect Score as the discriminant - helpful for data with homologous top hits, e.g. phospho or glyco (Tandem and Comet only)
☐ Use Gamma distribution to model the negatives (Tandem only)
☑ Force the fitting of the mixture model (bypass automatic mixture model checks)
☑ Use decoy hits to pin down the negative distribution.Decoy Protein names begin with: XXX_ (whitespace not allowed)
    ☑ Use Non-parametric model (can only be used with decoy option)
☐ Report decoy hits with a computed probability (based on the model learned).

Ignore charge states: ☑ +1 ☐ +2 ☐ +3 ☐ +4 ☐ +5

☐ Run ProteinProphet afterwards

Enter additional options to pass directly to the command-line (expert use only!) -c2.5 -PPM -eN

InterProphet Options [Show / Hide]

Figure 6.1. Screenshot showing the settings used for PeptideProphet refinement of the SUMO-remnant identifications.

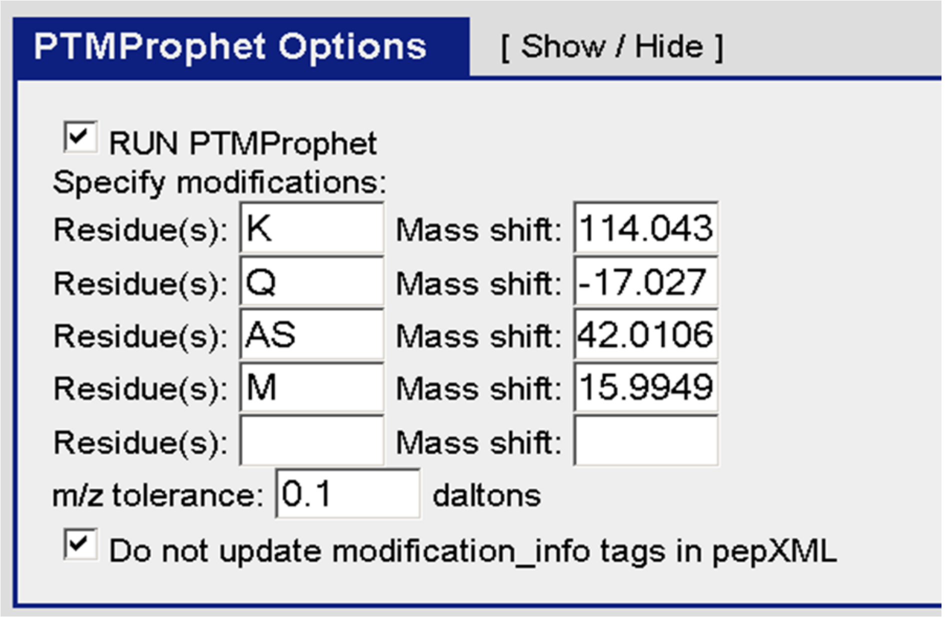
At the bottom under “Run Analysis,” click “Run XInteract!.” You will get one interact-filename.pep.xml file for each input file.

2. Run PTM Prophet

If you had four fractions that were all corresponding to the same sample, you will want to combine them at this step. Otherwise, just run PTM Prophet on the individual files. To combine the four fraction files, use iProphet (in TPP). Mouseover analysis Pipeline(Comet) chose Combine analyses. Upload your four files both .pep.xml and click do NOT use number of sibling searches (NSS) model and click Run Protein Prophet on these results. This will output one file that is the combined data, the file extension will be .ipro.pep.xml.

Prepare your peptide prophet output files to be run in PTM Prophet by editing the annotation for pyroglutamate. Open each one in notepad++, and replace the lines of text corresponding to peptide n-terminal pyroglutamate, *<modification_info mod_nterm_mass="-16.0187240729" modified_peptide="Q* with: *<modification_info modified_peptide="Q[111]* using the find and replace feature.

Now you can do the PTM Prophet analysis. Keep working in the same folder where all the files from the previous step are located. In TPP, click on “analyze PTMs” under “analysis pipeline” and analyze each file separately. Run PTMprophet to generate site-localization scores for SUMO-remnant modifications. For PTMprophet, the settings are pictured in Figure 6.2.



PTMProphet Options [Show / Hide]

☑ RUN PTMProphet

Specify modifications:

Residue(s):	K	Mass shift:	114.043
Residue(s):	Q	Mass shift:	-17.027
Residue(s):	AS	Mass shift:	42.0106
Residue(s):	M	Mass shift:	15.9949
Residue(s):		Mass shift:	

m/z tolerance: 0.1 daltons

☑ Do not update modification_info tags in pepXML

Figure 6.2. Screenshot showing the parameters used for PTMprophet. The output file will be in the same folder, named filename.ptm.pep. xml or .ptm.ipro.pep.xml (if you combined fractions beforehand).

3. Combine the heavy and light database search output

Theheavy and light lysine files are then combined using iProphet. Mouse-over analysis Pipeline(Comet) chose Combine analyses. The files have to be combined with the light first and then the heavy, so rename your heavy files with an x in front of them so that alphabetically the light will be read first. Also upload them light first and then heavy. Upload your two files both .ptm.pep.xml and select do NOT use number of sibling searches (NSS) model and click Run Protein Prophet on these results. Thiswill output one file that is the combined data for the heavy and light. The file extension is .ipro.pep.xml or .ipro.pep.ipro.xml (if you combined fractions beforehand). Check to be sure that the light file was read first by opening the resulting output file and search for “search_summary”. In the light file you’ll see only one entry for a modified K, in the heavy file you’ll see 2 entries for modified K one of which has massdiff=8.

4. Fix your PTM Prophet files for SILAC quantification

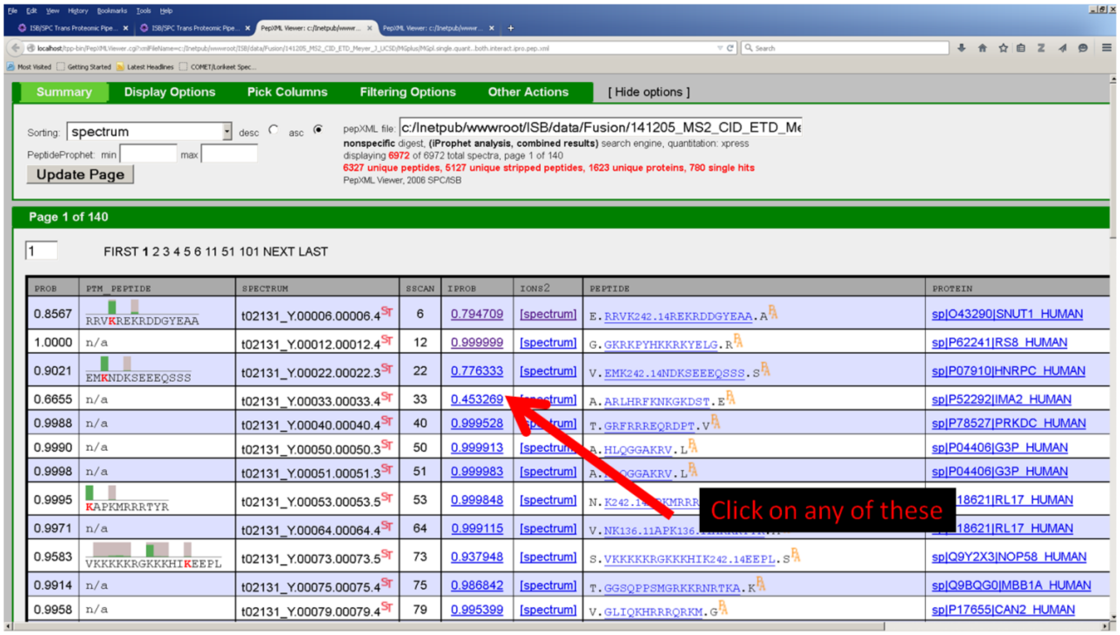
Currently TPP can’t run Xpress on our files, so we have to sendthe.ipro.pep.xml file to Jimmy Eng at TPP, so he can run a hacked version of Xpress. Upload your pep.xml AND corresponding mzXML files to google drive and email him the public link the files. Jimmy’s email is jke000@gmail.com, and you should mention you need the same quantification (quantification of the diGly lysines with heavy and light label). Open the resulting .ptm.interact.ipro.pep.xml file in TPP and click on “View?” PepXML to make sure that you can see the site localizations on the identified peptides.

5. Filer by minimum iProphet score to apply 1% FDR

Now the identifications can be filtered by minimum iProphet score. Upload the file (Files/browse files) in TPP then click on the View? PepXML link next to the file to see the data. Click on the iProb score of any peptide (Figure 6.3) (because the score table is the same for the entire file). Use the score table to determine what minimum iProb score corresponds to an error of 0.01 or less (Figure 6.4). Filter the file by going back to the viewer and click Filtering Options at the top. In the min iProphet probability, enter the score that corresponded to 0.01 (it will be something like 0.65 or 0.75 usually). Click also the exclude +1 charges (Figure 6.5). At the bottom, click Update Page. Once the page has updated, click on the Pick Columns tab at the top, click the “All” to add all the columns to the right. Click Update Page again. Click Other Actions choose Export Spreadsheet. The spreadsheet will be deposited in the same folder where you had the input file. Open it in excel, click Yes to accept the extension. Delete the columns that do not contain relevant information, including: retention time, compensation voltage, precursor intensity, collision energy, nss, nss_adj_prob, nrs, nrs_adj_prov, nse, nse_adj_prob, nsi, nsi_adj_prob, nsm, nsm_adj_prob. Keep the ions2 column even though it looks useless.

6. Filter the excel file to include only dglycyl-lysine peptides.

The excel file is then filtered to remove those peptides that are missing a quantification and also to make one table with only the diglycyl-lysine identifications and another with all the identifications (typically 10% of the peptides that are isolated


File Edit View History Bookmarks Tools Help
ISB/SPC Trans Proteomic Pipe...
ISB/SPC Trans Proteomic Pipe...
PepXML Viewer: c:/Inetpub/wwwr...
PepXML Viewer: c:/Inetpub/wwwr...
localhost/tpp-bin/PepXMLViewer.cgi?xmlFileName=c:/Inetpub/wwwroot/ISB/data/Fusion/141205_MS2_CID_ETD_Meyer_J_UCSD/MGplus/MGpl.single.quant..both.interact.ipro.pep.xml
Search
Most Visited
Getting Started
Latest Headlines
COMET/Lorikeet Spec...
Summary
Display Options
Pick Columns
Filtering Options
Other Actions
[Hide options]
Update Page
Click on any of these

Sorting: spectrum desc ○ asc ●

PeptideProphet: min max

pepXML file: c:/Inetpub/wwwroot/ISB/data/Fusion/141205_MS2_CID_ETD_Me

nonspecific digest, **(iProphet analysis, combined results)** search engine, quantitation: xpress

displaying 6972 of 6972 total spectra, page 1 of 140

6327 unique peptides, 5127 unique stripped peptides, 1623 unique proteins, 780 single hits

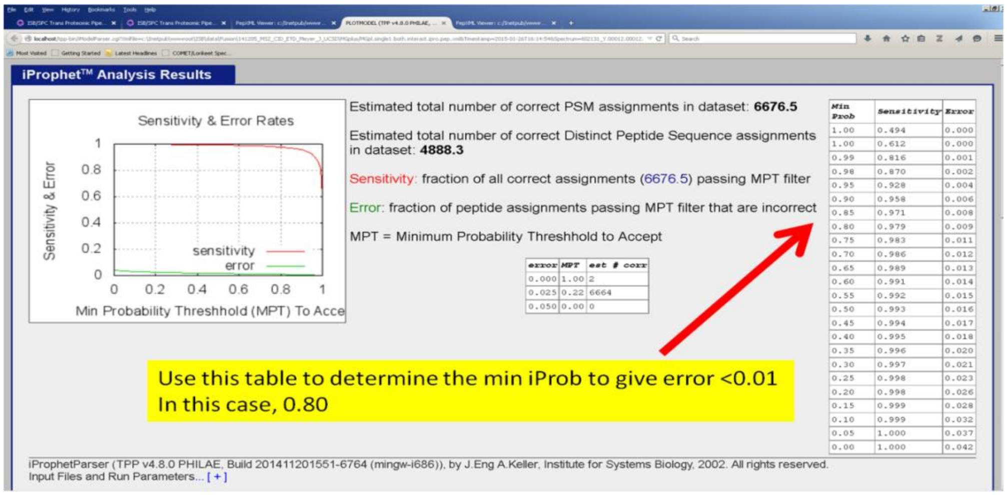
PepXML Viewer, 2006 SPC/ISB

Page 1 of 140

1 FIRST **1** 2 3 4 5 6 11 51 101 NEXT LAST

PROB	PTM_PEPTIDE	SPECTRUM	SSCAN	IPROB	IONS2	PEPTIDE	PROTEIN
0.8567	RRVKREKRDDGYEAA	t02131_Y.00006.00006.4	6	0.794709	[spectrum]	E.RRVK242.14REKRDDGYEAA.A	sp|O43290|SNUT1_HUMAN
1.0000	n/a	t02131_Y.00012.00012.4	12	0.999999	[spectrum]	G.GKRKPYHKKRKYELG.R	sp|P62241|RS8_HUMAN
0.9021	EMKNDKSEEEQSSS	t02131_Y.00022.00022.3	22	0.776333	[spectrum]	V.EMK242.14NDKSEEEQSSS.S	sp|P07910|HNRPC_HUMAN
0.6655	n/a	t02131_Y.00033.00033.4	33	0.453269	[spectrum]	A.ARLHRFKNKGKDST.E	sp|P52292|IMA2_HUMAN
0.9988	n/a	t02131_Y.00040.00040.4	40	0.999528	[spectrum]	T.GRFRRREQRDPT.V	sp|P78527|PRKDC_HUMAN
0.9990	n/a	t02131_Y.00050.00050.3	50	0.999913	[spectrum]	A.HLQGGAKRV.L	sp|P04406|G3P_HUMAN
0.9998	n/a	t02131_Y.00051.00051.3	51	0.999983	[spectrum]	A.[illegible]QGGAKRV.L	sp|P04406|G3P_HUMAN
0.9995	KAPKMRRRTYR	t02131_Y.00053.00053.5	53	0.999848	[spectrum]	N.K242.14[illegible]KMRRR[illegible]	[illegible]18621|RL17_HUMAN
0.9971	n/a	t02131_Y.00064.00064.4	64	0.999115	[spectrum]	V.NK136.11APK136.[illegible]	[illegible]18621|RL17_HUMAN
0.9583	VKKKKKRGKKKHIKEEPL	t02131_Y.00073.00073.5	73	0.937948	[spectrum]	S.VKKKKKRGKKKHIK242.14EEPL.S	sp|Q9Y2X3|NOP58_HUMAN
0.9914	n/a	t02131_Y.00075.00075.4	75	0.986842	[spectrum]	T.GGSQPPSMGRKKRNRTKA.K	sp|Q9BQG0|MBB1A_HUMAN
0.9958	n/a	t02131_Y.00079.00079.4	79	0.995399	[spectrum]	V.GLIQKHRRRQRKM.G	sp|P17655|CAN2_HUMAN

Figure 6.3. Click on any of the iProb scores to open the sensitivity/error plot.


File Edit View History Bookmarks Tools Help
PLOTMODEL (TPP v4.8.0 PHILAE, ...
PepXML Viewer: c:/Inetpub/www...
Most Visited
Getting Started
Latest Headlines
COMET/Lorikeet Spec...
Search
Sensitivity & Error Rates
Sensitivity & Error
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
sensitivity
error
Min Probability Threshhold (MPT) To Acce

iProphet™ Analysis Results

Estimated total number of correct PSM assignments in dataset: **6676.5**

Estimated total number of correct Distinct Peptide Sequence assignments in dataset: **4888.3**

Sensitivity: fraction of all correct assignments (6676.5) passing MPT filter

Error: fraction of peptide assignments passing MPT filter that are incorrect

MPT = Minimum Probability Threshhold to Accept

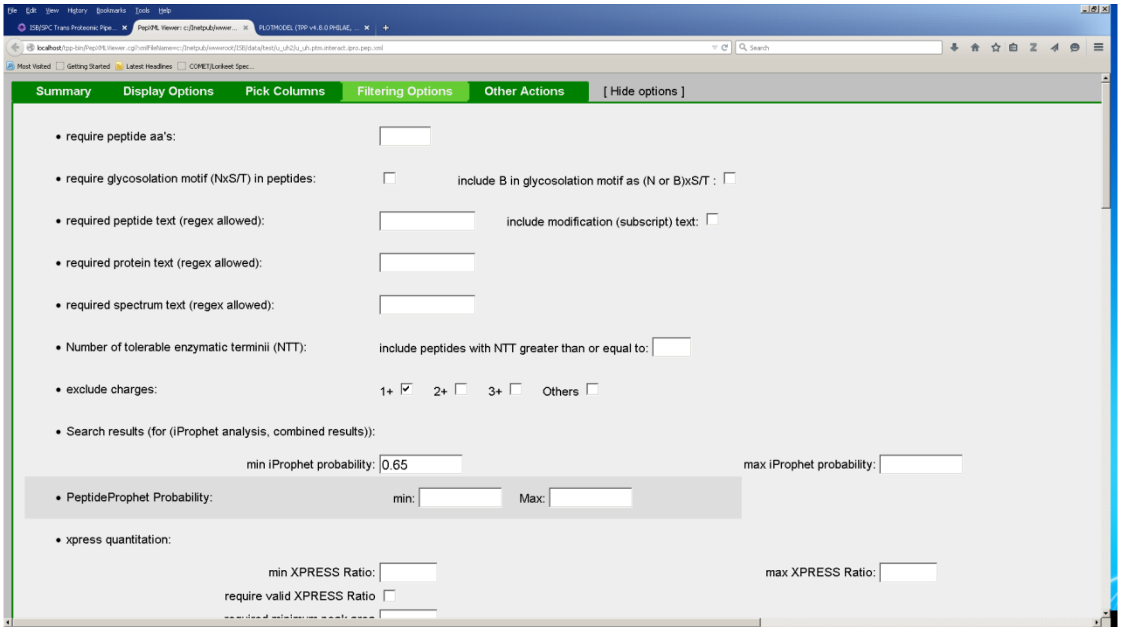
error	MPT	est # corr
0.000	1.00	2
0.025	0.22	6664
0.050	0.00	0

Min Prob	Sensitivity	Error
1.00	0.494	0.000
1.00	0.612	0.000
0.99	0.816	0.001
0.98	0.870	0.002
0.95	0.928	0.004
0.90	0.958	0.006
0.85	0.971	0.008
0.80	0.979	0.009
0.75	0.983	0.011
0.70	0.986	0.012
0.65	0.989	0.013
0.60	0.991	0.014
0.55	0.992	0.015
0.50	0.993	0.016
0.45	0.994	0.017
0.40	0.995	0.018
0.35	0.996	0.020
0.30	0.997	0.021
0.25	0.998	0.023
0.20	0.998	0.026
0.15	0.999	0.028
0.10	0.999	0.032
0.05	1.000	0.037
0.00	1.000	0.042

Use this table to determine the min iProb to give error <0.01
In this case, 0.80

iProphetParser (TPP v4.8.0 PHILAE, Build 201411201551-6764 (mingw-i686)), by J.Eng A.Keller, Institute for Systems Biology, 2002. All rights reserved.
Input Files and Run Parameters... [+]

Figure 6.4. Determine the appropriate value of iProbability to set for error < 0.01.


File Edit View History Bookmarks Tools Help
ISB/SPC Trans Proteomic Pipe...
PepXML Viewer: c:/Inetpub/wwwr...
PLOTMODEL (TPP v4.8.0 PHILAE, ...
localhost/tpp-bin/PepXMLViewer.cgi?xmlFileName=c:/Inetpub/wwwroot/ISB/data/test/u_uh2/u_uh.ptm.interact.ipro.pep.xml
Search
Most Visited
Getting Started
Latest Headlines
COMET/Lorikeet Spec...
Summary
Display Options
Pick Columns
Filtering Options
Other Actions
[Hide options]

- require peptide aa's:
- require glycosolation motif (NxS/T) in peptides: □ include B in glycosolation motif as (N or B)xS/T : □
- required peptide text (regex allowed): include modification (subscript) text: □
- required protein text (regex allowed):
- required spectrum text (regex allowed):
- Number of tolerable enzymatic terminii (NTT): include peptides with NTT greater than or equal to:
- exclude charges: 1+ ☑ 2+ □ 3+ □ Others □
- Search results (for (iProphet analysis, combined results)):

min iProphet probability: 0.65 max iProphet probability:

- PeptideProphet Probability: min: Max:
- xpress quantitation:

min XPRESS Ratio: max XPRESS Ratio:

require valid XPRESS Ratio □

Figure 6.5. Screenshot of the TPP window in which to set the minimum iProphet probability you determined from the statistics table.

from the IP will have GG-Ks). To remove the missing quantifications, filter the xpress column to remove the peptides with -1 and “Unavailable”. Save the resulting sheet as All. Then paste the following equation into the ions2 box (row 2) excel: “=OR(ISNUMBER(FIND(242, G2)),ISNUMBER(FIND(250,G2)))”. Copy that cell, select the whole row, and paste special formula to get the equation to run for all rows. This puts a FALSE if the GG-K is not verified and a TRUE if it is verified. The subset of verified diglycyl peptides can be selected out by sort data filter and deselect all the FALSE ones to get just the true ones. Save BOTH of the files both as an excel workbook and as a tab-delimited text file (which is used by R in the next steps).

C. Post-identification processing

1. Normalize distribution of quantification values, compute weighted average quantification values, and filter by localization score

Further filtering and processing of the search results is done with scripts written in R. Open Examples.R which sets the directory, reads the files, defines the class pepsum, etc and gives the order for running the functions. Open the all_functions.R and run all the lines in the all_functions.R. This defines all the functions you will need later. In the Examples.R, edit the working directory and run the line to see which files it thinks are the ones you want to analyze (all and digly) and put the file numbers into the brackets in the examples.R file. The R script has the option to output a tsv file, but right now that is set to FALSE because we won’t need it. Run the Examples.R script line by line. First, modification sites are filtered based on their site-localization score. Next, all unique protein sites are determined, and only those with quantification values are kept (in this case I used 0.75). A weighted average of quantification values is computed for protein sites identified by multiple peptides, with weights based on the integrated area. This information is output to a tsv file if the writetsv is set to T and not F. It is also saved as an object in R called mgpl.s.sig. This object is then used in further steps to correlate the sites between two different replicates, plot their quant values, and determine a linear fit. The weighted ratio is light/heavy (xpress score gives the raw ratio and “weighted”corresponds to the weighted average if the mod was in more than one peptide). The log ratio is light/heavy base 2. To get the heavy/light multiply the log.ratios by -1. Search the examples.R script for writetsv and name your file before running this line of code. At the end of the Examples.R script you can export a histogram of the weighted average ratios plotted against frequency. First, run all the function you need for the analysis to define them. The functions are in the file all_functions.R, which contains the following text:

```
#### this file contains all needed functions to go through SILAC quantification analysis
#### select all and run all within R
### create class for peptide identification data
setClass("pepsum",representation(summary="matrix",scans="ANY",specdir="character",
          specfiles="list",data="ANY", totalheat="matrix",residues="character",
          sequence="ANY",score="character",filetype="character",scanvec="ANY",
          modposition.protein="ANY",
chargevec="ANY",proteinmodlist="ANY",index="ANY",
          fraction="character",pepvec="ANY",provec="ANY",filenames="ANY",p1.i
ndex="ANY",
          modposition.peptide="ANY",countlist="list",pepraw="ANY",ionCoverage=
"list",
          ionCovSum="list",modindex="list",modsummary="list"))
### function to read peptide prophet results in tab-delim format
read.PepProph=function(input=files[2],type="PeptideProphet"){
          object<-new("pepsum")
          object@filetype="PeptideProphet"
          object@data<-read.delim(input,header=T)
          ### make scan and charge vectors
          scanlist<-strsplit(as.character(object@data$spectrum),split=".",fixed=T)
          scanlen<-length(scanlist)
          scanvec<-as.numeric(unlist(scanlist)[seq(2,length(scanlist)*4,by=4)])
          chargevec<-as.numeric(unlist(scanlist)[seq(4,length(scanlist)*4,by=4)])
          ### make peptide vector
          rawpepvec<-as.character(object@data$peptide)
          ### clean sequences into format for matchions
          #rawpepvec<-substr(rawpepvec,start=3, stop=nchar(rawpepvec)-2)
          peptempvec<-substr(rawpepvec,start=3, stop=nchar(rawpepvec)-2)
```

```
        peptempvec.l<-length(rawpepvec)
        peps<-rep(0,times=peptempvec.l)
        if(type=="PeptideProphet"){
            for(i in 1:peptempvec.l){
                ###match the string starting with [/[- followed by n integers and
"]"
                ### and replace with nothing
                peps[i]<-gsub("(\\[|\\[-)([0-9]+)(.)([0-9]+)(]+)", replacement="",
peptempvec[i])
                }
            for(i in 1:peptempvec.l){
                ###remove "n" from peptides with n-term acetyl
                if(unlist(strsplit(peps[i],split=""))[1]=="n"){
                    peps[i]<-substr(peps[i],start=2,stop=nchar(peps[i]))
                    }
                }
            }
        #### if format is inspect-style
        if(type=="inspect"){
            ### start by replacing those with nterminal pyroglu
            rawpepvec<-gsub(rawpepvec,pattern="(n)(\\[)(-
16.02)(\\])(Q)",replacement="Q[-17.027]")
            #### pyro glu done, move to the next mod
        rawpepvec<-
gsub(rawpepvec,pattern="(n)(\\[)(43.02)(\\])",replacement="[42.011]")
        ##### metox next
        rawpepvec<-
gsub(rawpepvec,pattern="(\\[)(147.04)(\\])",replacement="[15.995]")
        #### finally replace NEM-modified Cysteine
        rawpepvec<-
gsub(rawpepvec,pattern="(\\[)(228.06)(\\])",replacement="[125.048]")
        }
    if(type=="specnets"){
        ### start by replacing those with nterminal pyroglu
        rawpepvec<-gsub(rawpepvec,pattern="(n)(\\[)(-
16.02)(\\])(Q)",replacement="(Q,-17.027)")
        #### pyro glu done, move to the next mod
        nacetylindex<-grep(rawpepvec,pattern="(n)(\\[)(43.02)(\\])")
        nacetylstartres<-substr(rawpepvec[nacetylindex],start=9, stop=9)
        nacetyl_len<-length(rawpepvec[nacetylindex])
        for(i in 1:nacetyl_len){
            rawpepvec[nacetylindex][i]<-
paste("(",nacetylstartres[i],",+42.011)",rawpepvec[nacetylindex][i],sep="",collapse=""
)
            }
```

```
                rawpepvec<-gsub(rawpepvec,pattern="(n)(\\[)(43.02)(\\])([A-
Z]){1}",replacement="")
                ##### metox next
                rawpepvec<-
gsub(rawpepvec,pattern="(M)(\\[)(147.04)(\\])",replacement="(M,+15.995)")
                #### finally replace NEM-modified Cysteine
                rawpepvec<-
gsub(rawpepvec,pattern="(C)(\\[)(228.06)(\\])",replacement="(C,+125.048)")
                }
          if(type=="R"){
                ### start by replacing those with nterminal pyroglu
                rawpepvec<-gsub(rawpepvec,pattern="(n)(\\[)(-
16.02)(\\])(Q)",replacement="-17.027Q")
                #### pyro glu done, move to the next mod
                rawpepvec<-
gsub(rawpepvec,pattern="(n)(\\[)(43.02)(\\])",replacement="+42.011")
                ##### metox next
                rawpepvec<-
gsub(rawpepvec,pattern="(\\[)(147.04)(\\])",replacement="+15.995")
                #### finally replace NEM-modified Cysteine
                rawpepvec<-
gsub(rawpepvec,pattern="(\\[)(228.06)(\\])",replacement="+125.05")
                }
```

```
    object@pepvec<-peps
    object@scanvec<-scanvec
    object@chargevec<-chargevec
    return(object)
    }
#### function to filter identifications based on a minimum score
#### usage:
#### newobject<-removeLowLocalizations(object=[your pepsum object], minscore =
[your choice of localization cutoff score])
removeLowLocalization=function(object=mgpl.all,minscore=0.75){
    PTMscorelines<-as.character(object@data[,"ptm_peptide"])
    scores<-list()
    keepthese<-c()
    modposition<-list()
    line<-1
    for(i in 1:length(PTMscorelines)){
        tempscore<-
as.numeric(unlist(regmatches(PTMscorelines[i],gregexpr("[[:digit:]]+\\.*[[:digit:]]*",P
TMscorelines[i]))))
        if(length(which(tempscore>=minscore)>0)>0){
            #print("isnumeric")
            keepthese<-c(keepthese,i)
            n=which(tempscore>=minscore)
            modposition[[line]]<-
unlist(gregexpr("[[:digit:]]+\\.*[[:digit:]]*",PTMscorelines[i]))[which(tempscore>=minscore)]-((n-1)*7+2)
            line=line+1
            }
        }
    ### now have index of rows to keep
    ### and positions as a list of positions
    ### convert the list of peptide positions where length>1 into csv
    ### this works but the double mods are form of c(2,3)
    #test<-cbind(as.character(modposition),object@data[keepthese,])
    object@modposition.peptide<-modposition
    object@data<-object@data[keepthese,]
    object@pepvec<-object@pepvec[keepthese]
    return(object)
    }


#### function to determine the mod postion within a protein using the fasta DB
####  usage:
###     #### determine to position of each site ID in their protein
### mgpl.s.pos<-proteinPositions(object=mgpl.s,
###         fasta="C:/MSGFplus/database/110712_human.cc.fasta",
###         writetsv=FALSE,
###                name="mgplus.localized.tsv")
proteinPositions=function(fasta="C:/Users/JgMeyer/Documents/R/pepsum/pepsum/110712_human.cc.fasta",
    object=mgpl.single.pos,
    writetsv=FALSE,
    name="MG132.all.filtered.tsv"){
    ### read in fasta file
    require(seqinr)
    require(Biostrings)
    fastaobj<-read.fasta(fasta,seqtype="AA",as.string=TRUE)
    fastaacc<-substr(names(fastaobj),start=4,stop=9)
    datamat<-object@data
    #proteins<-levels(datamat[,"protein"])
    proteins<-as.character(unique(datamat[,"protein"]))
    uniqproteins<-substr(start=4,stop=9,proteins)
    allprotacc<-substr(as.character(datamat[,"protein"]),start=4,stop=9)
    ### replace object@pepvec with cleaned peptides
    peptempvec<-as.character(object@pepvec)
    peptempvec.l<-length(peptempvec)
    ### make empty vector for protein positions
    proteinposition<-rep(0,times=peptempvec.l)
    ####  loop through each line in the datamat
    proteinposition<-list()
```

```
        #### fix to correctly assign n-terminal position
        for(i in 1:peptempvec.l){
            print(i)
            #print(peptempvec[i])
            #print(datamat[i,"protein"])
            #currentpro<-substr(start=4,stop=9,datamat[i,"protein"])
            currentprot<-
fastaobj[[which(fastaacc==substr(start=4,stop=9,datamat[i,"protein"]))]][1]
            currentpep<-peptempvec[i]
            #print(currentpro)
            #print(currentpep)
            matched<-matchPattern(currentpep,currentprot)
            proteinposition[[i]]<-
matched@ranges[[1]][1]+(unlist(object@modposition.peptide[i])-1)
            }
        object@data<-
cbind(protein.position=as.character(proteinposition),object@data)
        #object@data[1,]
        object@modposition.protein<-proteinposition
        #name="testMGneg.tsv"
        if(writetsv==TRUE){
        write.table(object@data,file=name,quote=FALSE,sep="\t",row.names=F)
            }
```

```
		return(object)
		}
####	function to summarize and combine the unique protein site IDs
##		usage:
###		mgpl.s.ave<-summarizeProtPositions(object=mgpl.s.pos)
#mgpl.ave<-summarizeProtPositions()
#mgpl.ave@data[1,]
#mgpl.ave@modindex
### works with peptides containing multiple sites, sets their weighted ratio =0
summarizeProtPositions=function(object=mgpl.pos){
		proteinIDs<-unique(as.character(object@data[,"protein"]))
		prot.pos.list<-list()
		###  gives a list of proteins with their corresponding unique locations
		for(i in 1:length(proteinIDs)){
			#print(i)
			prot.pos.list[[proteinIDs[i]]]<-
unique(unlist(object@modposition.protein[which(object@data$protein==proteinIDs[i
])]))
			#print(proteinIDs[i])
			#print(prot.pos.list[[proteinIDs[i]]])
			}
		prot.lines.list<-list()
		#### gives the lines in object@data that correspond to each protein
```

```
    for(i in 1:length(proteinIDs)){
        #print(i)
        prot.lines.list[[proteinIDs[i]]]<-
which(object@data$protein==proteinIDs[i])
        }
    #range(unlist(prot.lines.list))
    position.index.list<-list()
    #### assign each unique position an index
    ### loop through the proteins
    index=1
    for(i in 1:length(proteinIDs)){
        #print(i)
        temp.positions<-
prot.pos.list[[which(names(prot.pos.list)==proteinIDs[i])]]

        temp.prot.lines<-which(object@data$protein==proteinIDs[i])  ###
gives row numbers of of mods in object@data

        protein.position.list<-object@modposition.protein[temp.prot.lines]  ###
gives the values of mod positions as vector
        ### set lines that have multiple mods to 0
        for(j in 1:length(protein.position.list)){
            #print(length(protein.position.list[[j]]))
            if(length(protein.position.list[[j]])>1){
                protein.position.list[[j]]<-0
                }
            }
        unique.positions<-unique(unlist(protein.position.list))
        unique.positions.l<-length(unique.positions)
        ### loop through the positions and assign those each an index number
        for(x in unique.positions){
            #print(x)
            ### if x!=0, do give an index
            if(x!=0){
                temprowlen<-
length(temp.prot.lines[which(protein.position.list==x)])
                for(j in 1:temprowlen){
                    #print(j)

     position.index.list[[temp.prot.lines[which(protein.position.list==x)][j]]]<-
index
                    }
                index=index+1
                }
            ###  if x==0 (peptide with two mods), assign index=0
            if(x==0){
```

```
                                        temprowlen<-
length(temp.prot.lines[which(protein.position.list==x)])
                                        for(j in 1:temprowlen){
                                                  #print(j)

            position.index.list[[temp.prot.lines[which(protein.position.list==x)][j]]]<-0
                                                  }
                                        }
                              }
                    }
                    ### which(position.index.list==111)
                    #### now loop through those values and take weighted averages of the
ones with more than 1 line
                    unique.indexes.len<-length(unique(unlist(position.index.list)))
                    unique.indexes<-unique(unlist(position.index.list))
                    #unique.indexes
                    weighted.ratios<-list()
                    linesum=0
                    for(j in 1:unique.indexes.len){
                              #print(j)
                              templines<-which(position.index.list==unique.indexes[[j]])
                              #print(templines)
                              object@data[templines,]
            if(unique.indexes[[j]]>=1){
                ### divide by temparea

     temparea=sum(object@data[templines,"light_area"])+sum(object@data[templines,"heavy_area"])
                templines.l<-length(templines)
                tempsum<-0
                linesum=linesum+templines.l
                weights<-rep(0,times=templines.l)
                for(i in 1:templines.l){
                    weights[i]<-(object@data[templines[i],"light_area"]+object@data[templines[i],"heavy_area"])/temparea
                    }
                for(i in 1:templines.l){

     tempsum=tempsum+object@data[templines[i],"xpress"]*weights[i]

     #print(object@data[templines[i],"light_area"]+object@data[templines[i],"heavy_area"])
                    #print(tempsum)
                    }
                for(i in 1:templines.l){
```

```
                weighted.ratios[[templines[i]]]<-tempsum
                }
            }
        if(unique.indexes[[j]]==0){
            templines.l<-length(templines)
            for(i in 1:templines.l){
                weighted.ratios[[templines[i]]]<-0
                }
            }
        }
    object@modsummary<-prot.pos.list
    object@modindex<-position.index.list
    object@data<-cbind(object@data,weighted.ratios=unlist(weighted.ratios))
    print("unique SUMO modified protein IDs")
    print(length(proteinIDs))
    print("unique SUMO modification sites from single-site peptides")
    print(length(index))
    length(prot.pos.list)
    ### write something to make these text and put them in one column
    #paste(unlist(prot.pos.list[1]))
    return(object)
    }
#### compute the weighted average of sites identified by with multiple sequences,
#### then determine which sites are outside at least one standard deviation
#### will produce a plot of quantification values as given in figure 3a
###          usage:
###          mgpl.s.sig<-getSignificant(allIDs=mgpl.s.all,targetIDs=mgpl.s.ave)
getSignificant=function(allIDs=mgpl.all,targetIDs=mgpl.ave,
             name="filename.filtered",
             writetsv=T,
             stdev=0.9404181){
        logRatiosAll<-log(allIDs@data$xpress,base=2)
        allIDs@data<-cbind(allIDs@data,logRatiosAll)
        meanAll<-mean(logRatiosAll)
        medianAll<-median(logRatiosAll)
        logRatiosAllnorm<-logRatiosAll-medianAll
        norm.median.all<-median(logRatiosAllnorm)
        ### get distribution of weighted averages for unique sites
        #targetIDs@modindex
        numuniq<-length(unique(unlist(targetIDs@modindex)))-1
        unique.position.singleline<-rep(0,times=numuniq)
        unique.position.singleweight<-rep(0,times=numuniq)
        #targetIDs@data$weighted.ratios
        for(i in 1:numuniq){
             unique.position.singleline[i]<-which(targetIDs@modindex==i)[1]
        unique.position.singleweight[i]<-
log(targetIDs@data[which(targetIDs@modindex==i)[1],"weighted.ratios"],base=2)
        }
    median(unique.position.singleweight)
    range(unique.position.singleweight)
    norm.uniq.pos.singleweight<-unique.position.singleweight-medianAll
    #2*sd(norm.uniq.pos.singleweight)
    if(length(stdev)==1){
        upperlim<-norm.median.all+2*stdev
        lowerlim<-norm.median.all-2*stdev
        onesdup<-norm.median.all+1*stdev
        onesddown<-norm.median.all-1*stdev
        }
    if(length(stdev)==0){
        upperlim<-norm.median.all+2*sd(logRatiosAllnorm)
        lowerlim<-norm.median.all-2*sd(logRatiosAllnorm)
        onesdup<-norm.median.all+1*sd(logRatiosAllnorm)
        onesddown<-norm.median.all-1*sd(logRatiosAllnorm)
        }
    #?pairlist
    #log(10)
    #unique.position.singleline[1]
    #norm.uniq.pos.singleweight[1]
```

```
names(norm.uniq.pos.singleweight)<-unique.position.singleline
sort.norm.uniq.pos.singleweight<-sort(norm.uniq.pos.singleweight*-1)
#median(sort.norm.uniq.pos.singleweight)
plot(x=1:numuniq,y=sort.norm.uniq.pos.singleweight)
abline(h=upperlim,col="red")
abline(h=lowerlim,col="red")
###
print("one standard deviation is")
print(onesdup)
abline(h=onesdup,col="red")
abline(h=onesddown,col="red")
#### now flag those IDs that are outside 2 std. dev.
idlen<-length(targetIDs@data$xpress)
oversigma<-rep(0,times=idlen)
count=0
#### loop through and test which are outside 2 std. dev
for(i in 1:numuniq){
        if(norm.uniq.pos.singleweight[i]>=onesdup |
norm.uniq.pos.singleweight[i]<=onesddown){
                #print(norm.uniq.pos.singleweight[i])
                count=count+1
                #print(count)
                oversigma[unique.position.singleline[i]]<-1
```

```
                        }
                }
        over2sigma<-rep(0,times=idlen)
        count=0
        #### loop through and test which are outside 2 std. dev
        for(i in 1:numuniq){
                if(norm.uniq.pos.singleweight[i]>=upperlim |
norm.uniq.pos.singleweight[i]<=lowerlim){
                        #print(norm.uniq.pos.singleweight[i])
                        count=count+1
                        #print(count)
                        oversigma[unique.position.singleline[i]]<-2
                        }
                }
        # for those missing their weighted average value, put in xpress value
        for(i in 1:length(targetIDs@data[,"weighted.ratios"])){
                if(targetIDs@data[i,"weighted.ratios"]==0){
                        targetIDs@data[i,"weighted.ratios"]<-
targetIDs@data[i,"xpress"]
                        }
                }
        targetIDs@data<-cbind(targetIDs@data,log.ratios=log(targetIDs@data[,"weighted.ratios"],base=2)-medianAll)
        targetIDs@data<-cbind(targetIDs@data,oversigma)
        #targetIDs@data<-cbind(targetIDs@data,over1sigma)
        ### part to output only unique significant changers
        #unique.position.singleline
        #targetIDs@data[,"log.ratios"]
        uniquelines<-targetIDs@data[unique.position.singleline,]
        changelines<-uniquelines[uniquelines[,"oversigma"]>=1,]
        nochangelines<-uniquelines[uniquelines[,"oversigma"]==0,]
        ### part to output only unique non changers
        ### now have new object with binary whether outside 2*sigma
        print(i)
        #### maybe write new table?
        print("unique sites")
        print(numuniq)
        print("over 2 stdev")
        print(length(oversigma[oversigma==2]))
        print("over 1 stdev")
        print(length(oversigma[oversigma==1]))
        if(writetsv){
        write.table(file=paste(name,".all.tsv",sep=""),targetIDs@data,quote=F,sep="\
t",col.names=T,row.names=F)
        write.table(file=paste(name,".changing.tsv",sep=""),changelines,quote=F,sep
="\t",col.names=T,row.names=F)
        write.table(file=paste(name,".nochange.tsv",sep=""),nochangelines,quote=F,
sep="\t",col.names=T,row.names=F)
            }
        return(list(allIDs,targetIDs,nochangelines,changelines))
        }
correlate.positions()
correlate.positions=function(object1=uv0.sig,object2=uv2.sig,name1="uv2hr",name2=
"uv8hr")
        {
        modsum1<-object1[[2]]@modsummary
        modsum2<-object2[[2]]@modsummary
        names(modsum1)<-substr(names(modsum1),start=4,stop=9)
        names(modsum2)<-substr(names(modsum2),start=4,stop=9)
        i=1
        matchvec<-c()
        for(x in names(modsum1)){
            print(which(names(modsum2)==x))
            if(length(which(names(modsum2)==x))>0){
```

```
            matchvec[i]<-x
            i=i+1
            }
        }
    #### loop through the match names and check their positions
    #### in the larger matrix
    #### take value of quant for each site put into a list of pairs
    quantpairs<-list()
    length(unlist(modsum1[matchvec]))
    modsum1.sub<-modsum1[matchvec]
    modsum2.sub<-modsum2[matchvec]
    uniprot<-substr(object2[[2]]@data[,"protein"],start=4,stop=9)
    object2[[2]]@data<-cbind(object2[[2]]@data,uniprot=uniprot)
    uniprot<-substr(object1[[2]]@data[,"protein"],start=4,stop=9)
    object1[[2]]@data<-cbind(object1[[2]]@data,uniprot=uniprot)
    hprp.values<-c()
    single.values<-c()
    for(x in names(modsum1.sub)){
        matched.positions<-
modsum2.sub[[x]][na.omit(match(modsum1.sub[[x]],modsum2.sub[[x]]))]
        for(y in matched.positions){
            obj2.protlines<-
object2[[2]]@data[which(as.character(object2[[2]]@data[,"uniprot"])==x),]
```

```
                    hprp.value<-
obj2.protlines[,"weighted.ratios"][obj2.protlines[,1]==y][1]
                    obj1.protlines<-
object1[[2]]@data[which(as.character(object1[[2]]@data[,"uniprot"])==x),]
                    single.value<-
obj1.protlines[,"weighted.ratios"][obj1.protlines[,1]==y][1]
                    if(is.na(single.value)==is.na(hprp.value)){
                            quantpairs[[paste(x,"_",y,sep="")]]<-
c(single.value,hprp.value)
                            hprp.values<-c(hprp.values,hprp.value)
                            single.values<-c(single.values,single.value)
                            }
                    }
              }
        par(cex.axis=1.5,cex=1.25)
        plot(log(hprp.values,base=2),log(single.values,base=2),ylim=c(),ylab=name1
,xlab=name2,main="correlation of SILAC quant values",pch=20)
        y=log(hprp.values,base=2)
        x=log(single.values,base=2)
        lm(x ~ y +0 ) ### not sure if +0 works
        abline(lm(x~y+0),col="red",lwd=2)
        abline(h=0)
        abline(v=0)
        print(summary(lm(x~y+0)))
        }
```

Here is the examples.R script:

```
#### before running any functions, they must be defined in your R workspace by running them
###             open all_functions.R, select all, and run within R
#### change working directory to locations of tab-delim output from TPP
setwd("C:/Inetpub/wwwroot/ISB/data/test/UVdata1")
#setwd("C:/Users/JGmeyer/my documents/R/silacquant")
### make object with list of files and then print them
files<-list.files()
files
### create class for peptide identification data
###
setClass("pepsum",representation(summary="matrix",scans="ANY",specdir="character",
        specfiles="list",data="ANY", totalheat="matrix",residues="character",
        sequence="ANY",score="character",filetype="character",scanvec="ANY",
        modposition.protein="ANY",
chargevec="ANY",proteinmodlist="ANY",index="ANY",
        fraction="character",pepvec="ANY",provec="ANY",filenames="ANY",p1.index="ANY",
        modposition.peptide="ANY",countlist="list",pepraw="ANY",ionCoverage=
"list",
        ionCovSum="list",modindex="list",modsummary="list"))
#### first read the peptide prophet results for SUMO-remnant-only IDs into pepsum
object
mgpl.s<-read.PepProph(input=files[2])
#### read all the IDs into pepsum object - used to get whole distribution for
normalization
mgpl.s.all<-read.PepProph(input=files[1])
#### remove those IDs with localization scores below an arbitrary value
mgpl.s<-removeLowLocalization(object=mgpl.s,minscore=0.75)
#### determine to position of each site ID in their protein
mgpl.s.pos<-proteinPositions(object=mgpl.s,
            fasta="F:/MSGFplus/database/110712_human.cc.fasta",
            writetsv=FALSE,
            name="mgplus.localized.tsv")
#### summarize and combine the unique protein site IDs
mgpl.s.ave<-summarizeProtPositions(object=mgpl.s.pos)
#### compute the weighted average of sites identified by with multiple sequences,
#### then determine which sites are outside at least one standard deviation
#### will produce a plot of quantification values as given in figure 3a
mgpl.s.sig<-getSignificant(allIDs=mgpl.s.all,targetIDs=mgpl.s.ave,
name="uv0.filtered",
                writetsv=T)
#### correlate the sites between two different replicates, plot their quant values, and
determine a linear fit
#### This is after running the previous code more than once to obtain several data
.sigs
newobject<-correlate.positions(object1=mgpl.s.sig,
          object2=uv8.sig,
          name1="anyname1",
          name2="anyname2")
##### additional examples below for UV data quant, normalizationm weighted
average
uv0<-read.PepProph(input=files[2])
uv0.all<-read.PepProph(input=files[1])
uv0<-removeLowLocalization(object=uv0,minscore=0.75)
uv0<-proteinPositions(object=uv0,
                fasta="F:/MSGFplus/database/110712_human.cc.fasta",
                writetsv=FALSE,
                name="mgplus.localized.tsv")
uv0<-summarizeProtPositions(object=uv0)
uv0.sig<-getSignificant(allIDs=uv0.all,
          targetIDs=uv0,
          writetsv=F,
          name="uv0.norm.sig.tsv")
```

```
uv2<-read.PepProph(input=files[9])
uv2.all<-read.PepProph(input=files[8])
uv2<-removeLowLocalization(object=uv2,minscore=0.75)
uv2<-proteinPositions(object=uv2,
            fasta="F:/MSGFplus/database/110712_human.cc.fasta",
            writetsv=F,
            name="UV2.localized.tsv")
uv2<-summarizeProtPositions(object=uv2)
uv2.sig<-
getSignificant(allIDs=uv2.all,targetIDs=uv2,writetsv=T,name="uv2.norm.sig.tsv")
#### correlate the sites between two different replicates, plot their quant values, and
determine a linear fit
newobject1<-correlate.positions(object1=uv0.sig, object2=uv2.sig)
uv8<-read.PepProph(input=files[13])
uv8.all<-read.PepProph(input=files[12])
uv8<-removeLowLocalization(object=uv8,minscore=0.75)
uv8<-proteinPositions(object=uv8,
            fasta="F:/MSGFplus/database/110712_human.cc.fasta",
            writetsv=F,
            name="UV8.localized.tsv")
uv8<-summarizeProtPositions(object=uv8)
uv8.sig<-
getSignificant(allIDs=uv8.all,targetIDs=uv8,writetsv=T,name="uv8.norm.sig.tsv")
```

```
#### show the correlation between UV2hr and UV8hr
newobject2<-correlate.positions(object1=uv2.sig,
object2=uv8.sig,name1="UV2hr",name2="UV8hr")
length(object@data[1,])
mgpl[[2]]@data[1,]
#### part to generate histograms as shown in figure 3a
###        set output object from getSignificant to: yourobject[[2]]->object
### 2 is always the number in the [[]] here, to get the Target IDs from the list of data
object<-uv0.sig[[2]]
?hist
#par(mfcol=c(2,1))
histtest<-hist(log(object@data[,"weighted.ratios"],2),breaks=200,main="weighted
average")
histtest
data.frame(histtest$breaks,histtest$counts)
write(histtest$counts,file="teest2.txt")
write(histtest$breakss,file="breaks2.txt")
### for sideways histogram either rotate the image from above or run barplot
barplot(horiz=T,histtest$counts)
##########################################################
#########   lines below not run
barplot(horiz=T,histtest$counts ,ylim=range)
?hist
```

```
?barplot
upperlim2sd<-mean(log(object@data[,26],2))+2*sd(log(object@data[,26],2))
lowerlim2sd<-mean(log(object@data[,26],2))-2*sd(log(object@data[,26],2))
upperlim1sd<-mean(log(object@data[,26],2))+1*sd(log(object@data[,26],2))
lowerlim1sd<-mean(log(object@data[,26],2))-1*sd(log(object@data[,26],2))
        abline(v=mean(log(object@data[,26],2)),col="blue")
        abline(v=upperlim2sd,col="red")
        abline(v=lowerlim2sd,col="red")
        abline(v=upperlim1sd,col="red")
        abline(v=lowerlim1sd,col="red")
hist(log(object@data$xpress,2),breaks=200,main="raw xpress scores")
upperlim2sd<-mean(log(object@data$xpress,2))+2*sd(log(object@data$xpress,2))
lowerlim2sd<-mean(log(object@data$xpress,2))-2*sd(log(object@data$xpress,2))
upperlim1sd<-mean(log(object@data$xpress,2))+1*sd(log(object@data$xpress,2))
lowerlim1sd<-mean(log(object@data$xpress,2))-1*sd(log(object@data$xpress,2))
        abline(v=mean(log(object@data$xpress,2)),col="blue")
        abline(v=upperlim2sd,col="red")
        abline(v=lowerlim2sd,col="red")
        abline(v=upperlim1sd,col="red")
```

2.Compare SUMO-site identifications with previously reported modifications

The overlap in identifications can be determined by comparing with other datasets that are available for lysine modifications. The first step of this process is to generate an object that R can use to compare to the datasets. This is done using the “filterfromall_short.R” script. For this script, you need the “All” tab-delimited file that you first generated in the the previous step (after you removed the empty columns but not any of the other data). Here is the text in the filterfromall.R script:

```
getwd()
setwd("C:/Inetpub/wwwroot/ISB/data/test/u_uh2")
files<-list.files()
files
uuh<-read.PepProph(input=files[10])
filterAll=function(object=mgpl.single.all,
        minscore=0.75,
        fasta="F:/MSGFplus/database/110712_human.cc.fasta",
        name="filename.tsv",writetsv=TRUE)
        {
        PTMscorelines<-as.character(object@data[,"ptm_peptide"])
        object1<-object
        #### object 1 only those with localization scores
        #which(PTMscorelines!="unavailable")
        object1@data<-object@data[PTMscorelines!="[unavailable]",]
        #nrow(object@data[PTMscorelines!="[unavailable]",])
```

```
object1@pepvec<-object@pepvec[PTMscorelines!="[unavailable]"]
#object1@pepvec[1]
#object1@data[1,]
### also filter based on 242\250 in peptide
peptides<-as.character(object1@data$peptide)
diglylines<-c(grep("242",peptides),grep("250",peptides))
object1@data<-object1@data[diglylines,]
#nrow(object1@data)
object1@pepvec<-object1@pepvec[diglylines]
object1@data$peptide
object1@pepvec
scorelines<-as.character(object1@data[,"ptm_peptide"])
scores<-list()
keepthese<-c()
modposition<-list()
line<-1
for(i in 1:length(scorelines)){
    tempscore<-as.numeric(unlist(regmatches(scorelines[i],gregexpr("[[:digit:]]+\\.*[[:digit:]]*",scorelines[i]))))
    if(length(which(tempscore>=minscore)>0)>0){
        #print("isnumeric")
        keepthese<-c(keepthese,i)
```

```
                        n=which(tempscore>=minscore)

                        modposition[[line]]<-
unlist(gregexpr("[[:digit:]]+\\.*[[:digit:]]*",scorelines[i]))[which(tempscore>=minscore)]-((n-1)*7+2)

                        line=line+1

                        }

                }

          ### now have index of rows to keep

          ### and positions as a list of positions

          ### convert the list of peptide positions where length>1 into csv

          object1@data<-object1@data[keepthese,]

          object1@data[,2]

          object1@modposition.peptide<-modposition

          object1@pepvec<-object1@pepvec[keepthese]

          #### now get protein positions and summarize

          ### read in fasta file

          require(seqinr)

          require(Biostrings)

          fastaobj<-read.fasta(fasta,seqtype="AA",as.string=TRUE)

          fastaacc<-substr(names(fastaobj),start=4,stop=9)

          datamat<-object1@data

          #proteins<-levels(datamat[,"protein"])

          proteins<-as.character(unique(datamat[,"protein"]))
```

```
        uniqproteins<-substr(start=4,stop=9,proteins)
        allprotacc<-substr(as.character(datamat[,"protein"]),start=4,stop=9)
        ### replace object@pepvec with cleaned peptides
        peptempvec<-as.character(object1@pepvec)
        peptempvec.l<-length(peptempvec)
        ### make empty vector for protein positions
        proteinposition<-rep(0,times=peptempvec.l)
        ####  loop through each line in the datamat
        proteinposition<-list()
        #### fix to correctly assign n-terminal position
        for(i in 1:peptempvec.l){
             print(i)
             currentprot<-
fastaobj[[which(fastaacc==substr(start=4,stop=9,datamat[i,"protein"]))]][1]
             currentpep<-peptempvec[i]
             #print(currentpro)
             #print(currentpep)
             matched<-matchPattern(currentpep,currentprot)
             proteinposition[[i]]<-
matched@ranges[[1]][1]+(unlist(object1@modposition.peptide[i])-1)
             }
        object1@data<-
cbind(protein.position=as.character(proteinposition),object1@data)
```

```
    #object@data[1,]
    object1@modposition.protein<-proteinposition
    #name="testMGneg.tsv"
    ###########################
    #### part to summarize and output unique positions
    proteinIDs<-unique(as.character(object1@data[,"protein"]))
    proteinIDs<-substr(proteinIDs,start=4,stop=9)
    prot.pos.list<-list()
    ###  gives a list of proteins with their corresponding unique locations
    for(i in 1:length(proteinIDs)){
        #print(i)
        prot.pos.list[[proteinIDs[i]]]<-unique(unlist(object1@modposition.protein[which(substr(start=4,stop=9,object1@data$protein)==proteinIDs[i])]))

    #unique(unlist(object1@modposition.protein[which(substr(start=4,stop=9,object1@data$protein)=="Q96T23")]))

        #prot.pos.list[["Q96T23"]]
        #print(proteinIDs[i])
        #print(prot.pos.list[[proteinIDs[i]]])
        }
    length(unlist(prot.pos.list))
```

```
	prot.lines.list<-list()
	#### gives the lines in object@data that correspond to each protein
	for(i in 1:length(proteinIDs)){
		#print(i)
		prot.lines.list[[proteinIDs[i]]]<-
which(substr(start=4,stop=9,as.character(object1@data$protein))==proteinIDs[i])
		}
	#range(unlist(prot.lines.list))
	position.index.list<-list()
	#### assign each unique position an index
	### loop through the proteins
	index=1
	index2=-1
	for(i in 1:length(proteinIDs)){
		print(i)
		temp.positions<-
prot.pos.list[[which(names(prot.pos.list)==proteinIDs[i])]]
		temp.prot.lines<-
which(substr(object1@data$protein,start=4,stop=9)==proteinIDs[i])  ### gives row
numbers of of mods in object@data
		protein.position.list<-object1@modposition.protein[temp.prot.lines]
### gives the values of mod positions as vector
		### set lines that have multiple mods to 0
```

```
        for(j in 1:length(protein.position.list)){

            #print(length(protein.position.list[[j]]))

            if(length(protein.position.list[[j]])>1){

                protein.position.list[[j]]<-0

                }

            }

        unique.positions<-unique(unlist(protein.position.list))

        unique.positions.l<-length(unique.positions)

        ### loop through the positions and assign those each an index number

        for(x in unique.positions){

            #print(x)

            ### if x!=0, do give an index

            if(x!=0){

                temprowlen<-

length(temp.prot.lines[which(protein.position.list==x)])

                for(j in 1:temprowlen){

                    #print(j)


     position.index.list[[temp.prot.lines[which(protein.position.list==x)][j]]]<-

index

                    }

                index=index+1

                }
            ###  if x==0 (peptide with two mods), assign negative index
            if(x==0){
                temprowlen<-
length(temp.prot.lines[which(protein.position.list==x)])
                for(j in 1:temprowlen){
                    #print(j)

     position.index.list[[temp.prot.lines[which(protein.position.list==x)][j]]]<-
index2
                    }
                index2=index2-1
                }
            }
        }
     object1@modsummary<-prot.pos.list
     object1@modindex<-position.index.list
     numuniq<-length(unique(unlist(object1@modindex)))
     uniq.indexes<-unique(unlist(object1@modindex))
     unique.position.singleline<-rep(0,times=numuniq)
     unique.position.singleweight<-rep(0,times=numuniq)
     #targetIDs@data$weighted.ratios
     i=1
     for(x in uniq.indexes){
        print(x)
        unique.position.singleline[i]<-which(object1@modindex==x)[1]
        i=i+1
        #unique.position.singleweight[i]<-
log(targetIDs@data[which(object1@modindex==i)[1],"weighted.ratios"],base=2)
        }
     uniquelines<-object1@data[unique.position.singleline,]
     object1@data<-uniquelines
     if(writetsv==TRUE){

     write.table(uniquelines,file=name,quote=FALSE,sep="\t",row.names=F)
        }
     return(object1)
     }
filterAll(object=uuh,name="uuh2.tsv")->uuh.filt
```

The filterfromR_short.R script generates an object and if you have set the writetsv to T it will also write a tsv output file. This object is then compared with data you have obtained from published source (which you also have to read into R objects). To make these comparisons, you first need to create objects in R from the published data. There are several small scripts to do this, they are called “readNature.R”, readPSP.R. Be sure to specify the directory where the data sets are found when you run these scripts. Then the sites are compared using the script “comparetables.R”.Scroll down to the bottom of comparetables.R to enter the name of object 1 and table (which needs to be in the format @data), which were output from filterfromall.R. Also name your output .tsv file.The output from this script is the full data table with extra columns to the right that have a + if the site was found in any of the compared datasets. Read the file nature.txt into an object in R using the function “readNature.R,” which contains the following text:

```
read.nature=function(input=files[1],type="nature",any=F,studies=T){
          object<-new("pepsum")
          object@filetype=type
          object@data<-read.delim(input,header=T)
          if(any){
                protacc<-levels(object@data[,1])
                for(x in protacc){
                        object@modsummary[[x]]<-
object@data[which(object@data[,1]==x),2]
                        }
                }
          if(studies){
                protacc<-levels(object@data[,1])
                study.lines<-
object@data[object@data[,"Count....Studies.SUMO.2.Modified."]>=1,]
                study.lines.l<-length(study.lines[,1])
```

```
        protacc<-levels(as.factor(as.character(study.lines[,1])))
        for(x in protacc){
                object@modsummary[[x]]<-
object@data[which(object@data[,1]==x),2]
                }
        }
    return(object)
    }
#object<-nature
read.nature(input=files[6])->nature
#nature@data[1,]
```

Then use the above function to actually read the nature tab-delimited text file with the following command:

```
read.nature(input=files[CHANGE THIS NUMBER TO THE NATURE.TXT FILE
NUMBER])->nature
```

This creates an object called nature that then you can use the comparetables script to compare to. Before you do this, you also want to create the PSP object so you can run all the comparisons at once. Use the readPSP.R and the four psp files (sumo, methyl, acetyl, Ub) to input all the data before running the comparetables.R. The “comparetables.R” script will generate a .tsv table that has extra columns for each of the dataset (SUMO, methyl, acetyl, Ub) and will enter a + whenever the site has been identified in any of these other data sets.Here is the comparetables.R script:

```
compare.tables=function(object1,
        table,
        natureobject=nature,
        ub=ub.psp,
        sumo=sumo.psp,
        methyl=methyl.psp,
        acetyl=acetyl.psp,
        write.table=T,
        write.extra.tables=F,
        name="test.tsv")
        {
        modsummary1<-object1@modsummary
        modsummary2<-natureobject@modsummary
        sumomods<-sumo.psp@modsummary
        ubmods<-ub.psp@modsummary
        methylmods<-methyl.psp@modsummary
        acetylmods<-acetyl.psp@modsummary
        #names(proteins1)<-proteins1
        proteins1<-names(modsummary1)
        proteins2<-names(modsummary2)
        proteins3<-substr(start=4,stop=9,table[,"protein"])
        sumo.proteins<-names(sumomods)
```

```
ub.proteins<-names(ubmods)

acetyl.proteins<-names(acetylmods)

methyl.proteins<-names(methylmods)

#proteins1<-substr(start=4,stop=9,proteins1)

#proteins2<-substr(start=4,stop=9,proteins2)

#names(modsummary1)<-proteins1

#names(modsummary2)<-proteins2

overlap.list<-list()

nooverlap.list<-list()

sumo.overlap<-list()

ub.overlap<-list()

count=0

for(x in proteins1){

    print(x)

    temppos<-which(x==proteins2)

    print(temppos)

    if(length(temppos)>0){

            #modsummary1[[x]]

            #length(modsummary1[[x]])

            #length(modsummary2[[x]])


if(length(na.omit(modsummary2[[x]][match(modsummary1[[x]],modsummary2[[x]])]))>0){
```

```
                overlap.list[[x]]<-
na.omit(modsummary2[[x]][match(modsummary1[[x]],modsummary2[[x]])])
                }
            #tempmatch<-match(modsummary1[[x]],modsummary2[[x]])
            #x<-"Q13547"

     if(length(na.omit(modsummary2[[x]][match(modsummary1[[x]],modsumma
ry2[[x]])]))==0){
                nooverlap.list[[x]]<-modsummary1[[x]]
                #names(x)
                }
            #print(tempmatch)
            count=count+1
            }
        }
     newcol<-rep("",times=length(object1@data[,1]))
     #### part to add a column with our study and insert pluses
     overlap.l<-length(overlap.list)
     for(j in 1:length(overlap.list)){
        tempprotname<-names(overlap.list[j])
        #unlist(x)
        print(j)
        templines<-which(natureobject@data[,"Protein"]==tempprotname)
```

```
                tempsites<-unlist(overlap.list[j])
                for(i in 1:length(tempsites)){

        newcol[templines[which(natureobject@data[templines,2]==tempsites[i])]]<-
"+"

        print(newcol[templines[which(natureobject@data[templines,2]==tempsites[i
])]])
                        }
                }

        #### part to add a column into our data table for presence in Nature
        tableproteins<-substr(start=4,stop=9,table[,"protein"])
        naturecol<-rep("",times=length(table[,1]))
        overlap.l<-length(overlap.list)
        table[1,]
        for(j in 1:length(overlap.list)){
                tempprotname<-names(overlap.list[j])
                #unlist(x)
                print(j)
                templines<-which(tableproteins==tempprotname)
                tempsites<-unlist(overlap.list[j])
                for(i in 1:length(tempsites)){
```

```
    naturecol[templines[which(table[templines,1]==tempsites[i])]]<-"+"

    print(naturecol[templines[which(natureobject@data[templines,2]==tempsites
[i])]])
            }
        }
    ### find what positions from the table overlap with the sumo psp sites and
return list
    count=0
    sumocol<-rep("",times=length(table[,1]))
    for(x in tableproteins){
        print(x)
        pos.in.sumo.modsummary<-which(x==sumo.proteins)
        pos.in.tableproteins<-which(x==tableproteins)
        print(temppos)
        tableproteins.sites<-table[pos.in.tableproteins,1]
        ### if the protein accession appears in the SUMO proteins
        ### check if the table
        if(length(pos.in.sumo.modsummary)>0){
            #modsummary1[[x]]
            #length(modsummary1[[x]])
            #length(modsummary2[[x]])
    if(length(na.omit(pos.in.tableproteins[match(tableproteins.sites,sumomods[[x
]])]))>0){
                ### this give the positions that should be + in the

    sumocol[na.omit(pos.in.tableproteins[match(tableproteins.sites,sumomods[[x
]])])]<-"+"
                }
            #tempmatch<-match(modsummary1[[x]],modsummary2[[x]])
            #x<-"Q13547"
            #print(tempmatch)
            count=count+1
            }
        }
    count=0
    ubcol<-rep("",times=length(table[,1]))
    for(x in tableproteins){
        print(x)
        pos.in.ub.modsummary<-which(x==ub.proteins)
        pos.in.tableproteins<-which(x==tableproteins)
        print(temppos)
        tableproteins.sites<-table[pos.in.tableproteins,1]
        ### if the protein accession appears in the SUMO proteins
```

```
            ### check if the table
            if(length(pos.in.ub.modsummary)>0){
                #modsummary1[[x]]
                #length(modsummary1[[x]])
                #length(modsummary2[[x]])

        if(length(na.omit(pos.in.tableproteins[match(tableproteins.sites,ubmods[[x]])
]))>0){
                    ### this give the positions that should be + in the

        ubcol[na.omit(pos.in.tableproteins[match(tableproteins.sites,ubmods[[x]])])]
<-"+"
                    }
                #tempmatch<-match(modsummary1[[x]],modsummary2[[x]])
                #x<-"Q13547"
                #print(tempmatch)
                count=count+1
                }
            }

        count=0
        methyl.col<-rep("",times=length(table[,1]))
        for(x in tableproteins){
```

```
        print(x)

        pos.in.methyl.modsummary<-which(x==methyl.proteins)

        pos.in.tableproteins<-which(x==tableproteins)

        print(temppos)

        tableproteins.sites<-table[pos.in.tableproteins,1]

        ### if the protein accession appears in the SUMO proteins

        ### check if the table

        if(length(pos.in.methyl.modsummary)>0){

            #modsummary1[[x]]

            #length(modsummary1[[x]])

            #length(modsummary2[[x]])


     if(length(na.omit(pos.in.tableproteins[match(tableproteins.sites,methylmods[
[x]])]))>0){

                ### this give the positions that should be + in the


     methyl.col[na.omit(pos.in.tableproteins[match(tableproteins.sites,methylmod
s[[x]])])]<-"+"

                }

            #tempmatch<-match(modsummary1[[x]],modsummary2[[x]])

            #x<-"Q13547"

            #print(tempmatch)

            count=count+1
```

```
                        }
                }
        count=0
        acetyl.col<-rep("",times=length(table[,1]))
        for(x in tableproteins){
                print(x)
                pos.in.acetyl.modsummary<-which(x==acetyl.proteins)
                pos.in.tableproteins<-which(x==tableproteins)
                print(temppos)
                tableproteins.sites<-table[pos.in.tableproteins,1]
                ### if the protein accession appears in the SUMO proteins
                ### check if the table
                if(length(pos.in.acetyl.modsummary)>0){

        if(length(na.omit(pos.in.tableproteins[match(tableproteins.sites,acetylmods[[
x]])]))>0){
                                        ### this give the positions that should be + in the

        acetyl.col[na.omit(pos.in.tableproteins[match(tableproteins.sites,acetylmods[
[x]])])]<-"+"
                                        }
                        count=count+1
                        }
```

```
		}
	output.table<-cbind(table,in.nature=naturecol,in.PSP.sumo=sumocol,in.PSP.ub=ubcol,in.PSP.acetyl=acetyl.col)
	alllist<-paste(substr(table[,"protein"],start=4,stop=9),table[,1],sep="_")
	naturelist<-paste(substr(table[naturecol=="+","protein"],start=4,stop=9),table[naturecol=="+",1],sep="_")
	sumolist<-paste(substr(table[sumocol=="+","protein"],start=4,stop=9),table[sumocol=="+",1],sep="_")
	ublist<-paste(substr(table[ubcol=="+","protein"],start=4,stop=9),table[ubcol=="+",1],sep="_")
	acetyllist<-paste(substr(table[acetyl.col=="+","protein"],start=4,stop=9),table[acetyl.col=="+",1],sep="_")
	if(write.extra.tables==T){

	write.table(naturelist,file="natlist.tsv",quote=F,sep="\t",col.names=T,row.names=F)
	write.table(sumolist,file="sumo.tsv",quote=F,sep="\t",col.names=T,row.names=F)

	write.table(ublist,file="ub.tsv",quote=F,sep="\t",col.names=T,row.names=F)

	write.table(acetyllist,file="acetyl.tsv",quote=F,sep="\t",col.names=T,row.names=F)

	write.table(alllist,file="alllist.tsv",quote=F,sep="\t",col.names=T,row.names=F)
		}
	if(write.table==T){

	write.table(output.table,file=name,quote=F,sep="\t",col.names=T,row.names=F)
		}
	return(output.table)
	}
test<-compare.tables(object1=uuh.filt,
	table=uuh.filt@data,
	natureobject=nature,
	ub=ub.psp,
        sumo=sumo.psp,

        methyl=methyl.psp,

        acetyl=acetyl.psp,

        write.table=T,

        write.extra.tables=F,

        name="uuh_comparetables.tsv")
```

3. Extract sequence windows for motif analysis

Motifs can next be extracted using the commands in the file “getmotif.R”, which uses the object you created from the filteredfromall as well. Scroll down to the bottom to specify which object (filename.filt) generated from the filtered from all script you want to run the Motif analysis on. Also specify the location of the fasta file.Here is the getMotif.R script:

```
#fasta="C:/MSGFplus/database/110712_human.cc.fasta"

#object<-mgpl.ave

getMotif=function(object,

        fasta="C:/MSGFplus/database/110712_human.cc.fasta",

        size=10)

        {

        datamat<-object@data

        require(seqinr)

        require(Biostrings)
```

```
fastaobj<-read.fasta(fasta,seqtype="AA",as.string=TRUE)

fastaacc<-substr(names(fastaobj),start=4,stop=9)

#proteins<-substr(as.character(datamat[,"protein"]),start=4,stop=9)

proteins<-as.character(datamat[,"protein"])

uniqproteins<-substr(start=4,stop=9,proteins)

allprotacc<-substr(as.character(datamat[,"protein"]),start=4,stop=9)

prot.l<-length(proteins)

hydrophobic<-c("F", "M", "P", "C", "L", "I", "W", "A", "V", "Q", "Y")

####

### retrieve the previous 15 residues and post 15 residues

### fix to add blanks for missing n-term or c-term values

sites.l<-length(unlist(object@modsummary))

window.vec<-rep(0,times=sites.l)

window.list<-list()

significant.vec<-rep(0,times=sites.l)

dbprotnames<-names(fastaobj)

targetprot.names<-names(object@modsummary)

#targetprot.names<-substr(start=4,stop=9,targetprot.names)

targetprot.l<-length(targetprot.names)


j=1

for(i in 1:targetprot.l){

        print(j)
        positions<-object@modsummary[[i]]+size
        tempprotname<-targetprot.names[i]
        seq<-fastaobj[[which(tempprotname==fastaacc)]][1]
        #### add "-" to each end of the sequence based on the max motif size
        seq<-paste(paste(rep("-",times=size),collapse=""),seq,paste(rep("-",times=size),collapse=""),sep="")
        seq.l<-nchar(seq)
        for(x in positions){
            window.vec[j]<-substr(seq,start=x-size,stop=x+size)
            if(nchar(window.vec[j])!=((size*2)+1)){print(i)}
            j=j+1
            #print(x)
            }
        }
    filtered<-unique(window.vec)
    filtered.l<-length(window.vec)
    de<-c("D","E")
    motifvec<-rep(0,times=filtered.l)
    change.vec<-rep(0,times=filtered.l)
    for(i in 1:filtered.l){
    ### assign 1 for known normal , 2 for inverted, 0 for other
        nextAA<-substr(filtered[i],start=size+2,stop=size+2)
        prevAA<-substr(filtered[i],start=size,stop=size)
        next.test<-which(hydrophobic==nextAA)
        prev.test<-which(hydrophobic==prevAA)
        filter
        if(length(next.test)>0){
        ### check number 4 for D/E
            if(length(which(substr(filtered[i],start=size-1,stop=size-1)==de)>=1)>0){
                print("inverted")
                motifvec[i]<-2
                }
            }
        if(length(prev.test)>0){
            ### check number 8 for D/E

    if(length(which(substr(filtered[i],start=size+3,stop=size+3)==de)>=1)>0){
                print("normal")
                motifvec[i]<-1
                }
            }
        }

    print("number of normal")
    print(length(which(motifvec==1)))
```

```
        print("number of inverted")
        print(length(which(motifvec==2)))
        print("number of new")
        print(length(which(motifvec==0)))
        na.omit(unique(window.vec[which(motifvec==1)]))
        na.omit(unique(window.vec[which(motifvec==2)]))
        na.omit(unique(window.vec[which(motifvec==0)]))
        newmotif<-na.omit(unique(filtered[which(motifvec==0)]))
        write(newmotif,file="newmotiflines.tsv")
        generalmotif<-na.omit(unique(filtered[which(motifvec==1)]))
        write(generalmotif,file="generalmotiflines.tsv")
        invertedmotif<-na.omit(unique(filtered[which(motifvec==2)]))
        write(invertedmotif,file="invertedmotiflines.tsv")
        allmotif<-na.omit(unique(filtered))
        write(allmotif,file="allmotiflines.tsv")

        }
getMotif(object=uuh.filt,fasta="F:/MSGFplus/database/110712_human.cc.fasta")
```

“Getmotif.R” will output four.tsv files with arbitrary-size sequence windows surrounding the modification site. These are the allmotif lines, the general motif lines, the inverted motif lines and the new motif lines. These sequence windows are used as input for IceLogo analysis.Open the ice logo program from the Downloads folder.

Open one of teh output.tsv files and click on the snowflake in the ice logo program and paste the text into the positive set window. In the right negative set window, choose human data and then click “make logo”. Click the tool icon at the bottom to make the start at -10 and you can also get the heat map.

Several edited versions of each script are included in the github repositories that have slight variations in functionality. Each file has a header describing the variations in their functionality.