NOVEL MASS SPECTROMETRY-BASED METHODS TO IDENTIFY AND QUANTIFY EXTRACELLULAR GLYCOPROTEINS

A Dissertation
Presented to
The Academic Faculty

by

Johanna Smeekens

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemistry & Biochemistry

Georgia Institute of Technology
May 2017

NOVEL MASS SPECTROMETRY-BASED METHODS TO IDENTIFY AND QUANTIFY EXTRACELLULAR GLYCOPROTEINS

Approved by:

Dr. Ronghu Wu, Advisor
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. Melissa Kemp
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Facundo Fernandez
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. Loren Williams
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. M.G. Finn
School of Chemistry & Biochemistry
Georgia Institute of Technology

Date Approved: March 31, 2017

ACKNOWLEDGEMENTS

I certainly would not have made it this far without the support of many wonderful people. First, I would like to thank my parents for their unwavering love, support and encouragement over the last five years. You believed in me through every hurdle and obstacle I faced, and I truly couldn't have made it this far without your support. I'm grateful to have inherited your work ethic and the will to never give up. I'd also like to thank my sisters, Bridget and Amelia for their love and encouragement...and endless bitmojis, memes, and snapchats to make me laugh when I needed it.

Many faculty at Georgia Tech have been incredibly helpful throughout this process, especially my advisor, Ronghu Wu. Thank you for your mentorship, guidance, and commitment to my success. Joining a brand new group was an invaluable experience that afforded many opportunities to grow in research and leadership, and I am grateful for everything we have accomplished together. I also appreciate the guidance and support from Dr. Facundo Fernandez and Dr. Larry Bottomley throughout the many stages of graduate school. Thanks to my committee, Dr. M.G. Finn, Dr. Facundo Fernandez, Dr. Melissa Kemp, Dr. Loren Williams, and Dr. Ronghu Wu, for their advisement and assistance in completing this PhD. A special thanks also goes to Dr. Christy O'Mahony, for her endless guidance, willingness to listen, and for making teaching so enjoyable.

I would like to thank current and former lab mates in the Wu group for making the day-to-day of graduate school more enjoyable, especially Haopeng Xiao and Jay Suttapitugsakul for the many laughs and meals together. A special thanks goes to Dr. Weixuan Chen, who has been a huge support since day one when we walked into a brand

new lab together. His guidance in research, graduate school, and life have been incredibly helpful and essential for my success.

I am grateful for the many amazing friends I've made at Georgia Tech, especially Alyson Colin, Laura Dinger, Hailey Bureau, Allison Geoghan, Alyssa Blake and Erin Gawron. Thanks for the many laughs, conversations and support over the years. I would like to especially thank the students who came before me, Rachel Stryffeler, Megan Mann and Kim Clarke, for their guidance throughout every step of graduate school (including writing this dissertation) – I truly would have been lost without them. I am also grateful for meeting many successful female graduate students through Women in Chemistry. Being a part of the group has been essential to my success in graduate school.

I am especially thankful for re-meeting Brittany McCormick, who has been an incredible friend and support the past four years. I am so grateful we had a second chance at becoming friends after we missed the opportunity at Michigan State. I don't know how I would have survived without your friendship the past four years! I am also especially thankful for Christina Richardson and Liz Ostergaard, who have been amazing friends since high school. Thanks for all of the encouraging texts, funny stories and inspirational quotes to keep me going.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

#	N-glycosylation site
*	Heavy lysine
ACN	Acetonitrile
AGC	Automatic gain control
Arg-C	Endoproteinase Arg-C
Asn	Asparagine
Asp	Aspartic acid
Asp-N	Endoproteinase Asp-N
BMP-2	Bone morphogenetic protein-2
CD	Cluster of differentiation
CNBr	Cyanogen bromide
CuAAC	Copper-catalyzed azide-alkyne cycloaddition
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DBCO	Dibenzocyclooctyne
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EMT	Epithelial-mesenchymal transition
ER	Endoplasmic reticulum
ES	Expansion segment
ESI	Electrospray ionization

FA	Formic acid
FBS	Fetal bovine serum
FDR	False discovery rate
GalNAc	N-acetylgalactosamine
GalNAz	N-azidoacetylgalactosamine
GlcNAc	N-acetylglucosamine
Glu-C	Endoproteinase Glu-C
GPI	Glycophosphatidylinositol
HBP	Hexosamine biosynthesis pathway
HEK 293T	Human embryonic kidney cells 293
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HILIC	Hydrophilic interaction liquid chromatography
ICAT	Isotope-coded affinity tags
iTRAQ	Isobaric tags for relative and absolute quantification
LC	Liquid chromatography
LDA	Linear discriminant analysis
LT	Lys-C and trypsin digestion
Lys-C	Lysyl endopeptidase
m/z	Mass-to-charge
MALDI	Matrix-assisted laser desorption ionization
MEGM	Mammary epithelial cell growth medium
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS^2	Tandem mass spectrum
NG	NTCB, Glu-C digestion

NLT	NTCB, Lys-C and trypsin digestion
NTCB	2-nitro-5-thiocyanobenzoic acid
PBS	Phosphate-buffered saline
PNGase F	Peptide-N-glycosidase F
PSA	Prostate-specific antigen
rRNA	Ribosomal RNA
SDC	Sodium deoxycholate
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SILAC	Stable isotope labeling with amino acids in cell culture
SNARE	Soluble *N*-ethylmaleimide-sensitive fusion protein attachment protein receptor
TFA	Trifluoroacetic acid
TGF-β	Transforming growth factor-β
TMT	Tandem mass tag
UPS	Ubiquitin-proteasome system

SUMMARY

Glycosylation is one of the most common and essential protein modifications in cells. This modification plays critical roles in protein folding, trafficking and stability, and regulates many cellular events. Abnormal glycosylation is often correlated with diseases such as cancer and infectious diseases. The vast majority of extracellular proteins are glycosylated; these cell surface and secreted glycoproteins are crucial for extracellular interactions, cell signaling and immune response, and also reflect the disease and developmental status of the cell. They serve as excellent therapeutic targets due to their easy accessibility by small molecules and macromolecules. Secreted glycoproteins are also a promising non-invasive source of diagnostic biomarkers, since they exist in bodily fluids and can be easily accessed for disease diagnosis and monitoring.

Due to their critical importance in cellular processes and relevance in disease, extracellular glycoproteins contain a wealth of information regarding the molecular mechanisms of disease. Comprehensive analysis of these glycoproteins will provide the foundation to identify diagnostic biomarkers and therapeutic targets. Modern mass spectrometry (MS)-based proteomics enables comprehensive and site-specific analysis of protein modifications, including glycosylation. However, global analysis of glycoproteins is extremely challenging. Many glycoproteins are present at low abundance in complex biological samples, and therefore require effective separation prior to MS-based analysis. Furthermore, glycans are highly heterogeneous, and therefore the mass tag necessary to pinpoint glycosylation sites is largely unpredictable.

The work in this thesis focuses on the development of mass spectrometry-based methods to globally analyze cell surface and secreted glycoproteins. Chapter 1 provides a brief overview of MS-based proteomics and glycoproteomics analysis, including current methods and challenges. Chapter 2 describes the novel method developed for site-specific analysis of the cell surface N-glycoproteome. Chapter 3 illustrates an application of this method to investigate cell surface glycoprotein changes throughout the epithelial-mesenchymal transition. Chapter 4 focuses on the comprehensive analysis of secreted proteins and glycoproteins from *Saccharomyces cerevisiae*, which is an excellent model system for eukaryotic cells. Finally, Chapter 5 provides an enhanced digestion method for improved membrane protein identification by MS. Altogether, this work affords new opportunities and methods for large-scale analysis of extracellular glycoproteins, which can be extensively applied to further decode the extracellular glycoproteome.

CHAPTER 1. INTRODUCTION TO MASS SPECTROMETRY-BASED PROTEOMICS AND GLYCOPROTEOMICS

1.1 Mass spectrometry-based proteomics

Proteins are one of the most abundant macromolecules within a cell, comprising about 50% of the total dry mass, and play key roles in all cellular processes and functions.[1] As a result, a comprehensive understanding of cellular biology and all related processes require knowledge about protein regulation, interactions, function and expression. With the advent of soft ionization sources, especially electrospray ionization (ESI)[2] and matrix-assisted laser desorption/ionization (MALDI)[3], mass spectrometry (MS) has become an increasingly powerful tool to investigate bio-macromolecules.[4-9] MS-based proteomics techniques offer a unique opportunity to systematically study proteins, which is beyond the reach of conventional biochemical methods.[10-11] Classical antibody-based protein identification methods depend entirely on the availability and quality of antibodies. In addition, antibodies are expensive and the experimental procedures are typically time-consuming, labor-intensive, and relatively low-throughput. Alternatively, MS-based techniques can assist in the confident global identification and quantification of proteins without the use of antibodies, due to their high-throughput nature.[12-14]

Many initial protein studies analyzed selected purified proteins to investigate their structure, function or interactions. Now, there is greater interest in comprehensively studying all proteins, or the proteome, of a biological sample. Understanding the whole

proteome and its structure, function and dynamics has been a central focus of protein biology.[15] Protein abundances range from several copies to millions of copies per cell, and are extremely dynamic in response to internal and external stimuli, which makes their comprehensive analysis quite complex.[16] MS-based methods provide a platform to perform comprehensive protein analysis, through three main approaches: top-down, middle-down and bottom-up proteomics.

1.1.1 Top-down proteomics

Top-down proteomics approaches analyze intact proteins with mass spectrometry, in hopes of gaining complete protein characterization. Protein ions are fragmented in the gas phase, and both intact parent and fragment masses are used to determine protein sequences.[17] Complete coverage at both the intact and fragment ion levels is ideal for comprehensive analysis, often times resulting in higher sequence coverage compared to bottom-up proteomics.[18-19] In theory, analyzing intact proteins makes the entire sequence available, thereby enabling the identification of specific protein isoforms and potentially improving protein quantification, since protein abundances are used instead of individual peptides.[20-22] However, there are still technical challenges in this field, including the lack of effective fractionation methods for intact protein separation that are compatible with MS/MS analysis.[23]

1.1.2 Middle-down proteomics

Between analyzing intact proteins with top-down workflows and smaller peptides with bottom-up workflows exists an approach that exploits strengths from both methods into middle-down proteomics. With this approach, proteins are digested into medium-

sized peptides (3 – 15 kDa) which provide more complete sequence information than bottom-up methods, and can be more easily fractionated prior to MS analysis by many common liquid chromatography (LC) techniques, compared to top-down methods.[24] Here, longer peptides are more likely to have unique sequences and correlate more specifically to protein isoforms, compared to bottom-up methods. Since the most commonly used protease, trypsin, generates shorter peptides (0.5 – 3 kDa), alternate proteases such as Lys-C, Asp-N, Arg-C, or Glu-C must be employed to generate these longer peptides.[25-27] Middle-down techniques have been successfully applied to a variety of protein studies, including histones, polyubiquitin chains and ribosomal proteins.[28-31] However, middle-down proteomics is still less widespread than top-down or bottom-up approaches.[29]

1.1.3 Bottom-up proteomics

Bottom-up proteomics, also called 'shotgun proteomics,' is the most common technique for global protein analysis. These methods rely on protein digestion to generate thousands of peptides which are subsequently separated via LC before analysis with tandem mass spectrometry (MS/MS).[32] Large-scale data analysis involves database searching to match acquired tandem mass spectra (MS^2) with theoretical spectra and determine the corresponding amino acid sequence of the peptide.[33-35] Protein characterization is subsequently based on peptides identified in the sample.[36] Contrary to top-down proteomics, fractionation of peptides before MS analysis is relatively easy to attain. Due to sample complexity, often times multidimensional LC is employed, to obtain ideal separation before analysis with MS.[32]

Depending on the protease used, peptide identifications are somewhat limited based on the frequency of cleavage sites within a protein sequence, and the corresponding bias of mass spectrometry for relatively short peptides (about 10-20 amino acids in length). Trypsin has become the gold standard for large-scale digestion, which specifically cleaves at the C-termini of lysine and arginine residues and generates peptides within the preferred mass range and with a basic residue at the C-termini. This approach has been widely successful for large scale proteome analysis, and is at the core of proteomics achievements.[32, 37-38] Currently, bottom-up proteomics is the main mode of comprehensive MS-based proteome analysis, and is the approach used in this thesis work.

1.1.4 Quantitative proteomics

Another main focus of large-scale protein analysis is quantification of the proteome. Many studies are interested in protein expression changes in response to a stimulus, *i.e.* drug treatment, pathway inhibition, altered growth conditions, etc. MS-based methods are not inherently quantitative, due to many variables including sample handling, varying ionization efficiency between peptides, and differing retention on chromatographic columns.[39] However, these methods can be coupled with isotope labeling to obtain relative protein abundances. Peptides with the same chemical composition, only differing by isotopes at specific positions, are expected to behave the same way during chromatographic separation and mass spectrometric analysis. Therefore, the peptides will have the same retention time, but can be differentiated through extracted ion chromatograms based on their mass-to-charge (m/z) ratios.[12]

1.1.4.1 Metabolic isotope labeling

Isotopes can be incorporated into media during cell growth with individual elements (^{15}N) or amino acids (Leu, Arg, Lys, Tyr) containing ^{13}C and/or ^{15}N.[40] For example, stable isotope labeling with amino acids in cell culture (SILAC) incorporates heavy lysine and arginine residues into proteins, which is particularly useful for labeling tryptic digestion peptides.[39, 41] In traditional SILAC studies, two cell culture flasks are used: one culture is labeled with 'heavy' amino acids and typically treated with a stimulus, while the other culture is labeled with normal (or 'light') amino acids and grown under standard conditions. For these approaches, heavy and light peptide pairs in the full mass spectrum are compared to determine the relative abundance ratio for each peptide, which is then inferred to determine protein abundance changes.[12] These methods are limited to cell culture and animal models, and cannot be applied to tissue or blood samples.

1.1.4.2 Chemical isotope labeling

Chemical derivatization is also employed to introduce isotopes by targeting specific functional groups within peptides. Tandem mass tags (TMT) and isobaric tags for relative and absolute quantification (iTRAQ) labeling methods both target free amines. Briefly, these isobaric tags react with free amines in peptides, resulting in identical mass tags on each peptide and identical behavior during chromatographic separation. When peptides are fragmented for MS/MS, these isobaric tags produce fragments with different masses, which are then compared to obtain the relative abundance changes of the peptide. TMT and iTRAQ are both multiplexed methods, capable of quantifying multiple samples at once: iTRAQ offers 4 or 8 plex, and TMT offers 2, 6 or 10 plex kits.[42-43] Multiplexed techniques have advantages compared to duplex techniques since they can

simultaneously quantify protein expression in up to 10 samples, which is particularly useful when quantifying proteins in many samples, or at many time points of a reaction or drug treatment.

Another common chemical isotope labeling method uses isotope-coded affinity tags (ICAT), which contain a biotin handle, and label free cysteines within proteins. Two protein samples can be quantified simultaneously: one is labeled with a heavy isotope tag, and the other with a light tag. After digestion, isotopically labeled peptides are separated with avidin affinity chromatography and analyzed with LC-MS/MS. Similar to SILAC, heavy and light peptide pairs are quantified based on their relative abundances in the full mass spectrum.[44-45]

1.2 Glycoproteomics

Protein modifications are essential in biological systems and involved in nearly every cellular event.[46-50] There has been great interest in the comprehensive and quantitative analysis of modified proteins to obtain a better understanding of protein functions and disease mechanisms, which could lead to the discovery of effective biomarkers and drug targets.[51-53] Although MS-based proteomics has been applied to globally identify modified proteins, pinpoint the modification sites, and quantify their abundance changes[54-58], it is still extraordinarily challenging to achieve these goals in complex biological samples for multiple reasons.[59-60]

Protein N-glycosylation is one of the most common and important modifications; it frequently initiates cell signal transduction and regulates cell-cell communication and cell-matrix interactions.[61-62] Based on predictions and computational results, about half of

all mammalian proteins are glycosylated at any given time.[63] During this modification, a sugar chain, or glycan, is attached to a protein through either asparagine (N-glycosylation), serine/threonine (O-glycosylation), or tryptophan (C-glycosylation) and is further modified by a variety of enzymes. The work presented here focuses only on N-glycosylation, which occurs on the consensus motif N-X-S/T, where X is any amino acid except proline.[64] This is a non-template driven process, which thereby generates a complex variety of glycans. Glycosylation often regulates protein folding, stability and trafficking, and is also required for many cellular processes, including extracellular interactions and immune response.[65-66] Aberrant glycosylation is often correlated with human diseases including cancer and infectious diseases.[53, 67-68] Due to the critical importance, frequency and biological implications of glycosylation, there has been increased focus on global analysis of protein glycosylation.

1.2.1 Extracellular glycoproteins

Extracellular proteins play critical roles in cellular interactions, and the vast majority, if not all, are glycosylated. Glycoproteins on the cell surface and in the secretome dictate extracellular interactions, cell signaling and immune response, and also reflect the disease and developmental status of the cell.[69-70] They serve as excellent therapeutic targets due to their easy accessibility by small molecules and macromolecules.[71-72] The majority of FDA approved drugs target membrane or secreted proteins.[73-74] These glycoproteins also serve as a promising non-invasive source of diagnostic biomarkers, since secreted proteins can be found in plasma/blood samples and other bodily fluids which are easily accessed for diagnosis and monitoring purposes.[75] Several secreted glycoprotein biomarkers have been verified by the FDA, including

prostate-specific antigen (PSA) for prostate cancer, CA125 for ovarian cancer, and HER2/NEU for breast cancer.[76-77] Comprehensive analysis of these extracellular glycoproteins will have significant impacts in biomedical and disease research.[77] However, comprehensive identification of cell surface glycoproteins is even more challenging compared to intracellular proteins. It requires specific separation prior to MS, which magnifies the difficulties associated with modified proteins mentioned above.

1.2.2 Difficulties with analysis

Modified proteins are present at much lower abundance within a cell, compared to many non-modified proteins. Previous studies have shown that enzymes responsible for glycosylation and other post-translational modifications tend to have less than 500 copies per cell. Contrarily, proteins involved in protein synthesis (folding, translation, etc.) are present at much higher abundances, greater than 100,000 copies per cell.[78] When MS analysis is performed on complex biological samples containing low and high abundance proteins, low abundance glycoproteins are masked by highly abundant proteins, and are rarely detected.[79] For this reason, the effective and comprehensive identification of glycoproteins requires enrichment methods capable of separating glycoproteins from non-glycosylated proteins.

Glycosylation is a non-template driven process, resulting in a hypothetically infinite variety of glycan structures. As such, the structure of a glycan at a specific site can vary greatly for the same protein. Furthermore, the number of glycosylation sites occupied within a protein also differs. This micro- and macro-heterogeneity makes MS-based global analysis of glycoproteins more challenging, since the glycan mass tag and

resulting shift are largely unknown.[80] Successful enrichment is difficult due to glycan structure variability.[81] Several enrichment methods are discussed below, but each has disadvantages that hinder comprehensive glycoprotein analysis.

Analyzing extracellular glycoproteins specifically, which is the focus of this thesis, further complicates enrichment and identification methods. The majority of extracellular proteins are predicted to be glycosylated, yet their low abundance and difficulties in separation prevent their comprehensive analysis. In addition to extracellular glycoproteins being masked by higher abundance non-modified proteins during MS analysis, they also are masked by intracellular glycoproteins. Therefore, there is a demand for effective methods that specifically target cell surface and extracellular glycoproteins for MS-based proteomics.

1.2.3 Current methods

Due to the growing recognition of glycoprotein significance in disease and biological processes, several enrichment methods have been developed for large-scale glycoprotein analysis. In general, enrichment methods take advantage of the common properties of glycans, including their many hydroxyl groups and hydrophilicity. While considerable progress has been made in large-scale glycoprotein analysis, it is still extremely challenging to attain complete glycoproteome analysis. The most common techniques utilize hydrophilic interaction liquid chromatography, lectins, boronic acid or hydrazide chemistry for enrichment, but the majority are not appropriate for extracellular glycoprotein enrichment.

1.2.3.1 Hydrophilic interaction liquid chromatography

Although glycan structures are very heterogeneous, all glycopeptides contain hydrophilic carbohydrate groups, which are exploited for glycopeptide enrichment with hydrophilic interaction liquid chromatography (HILIC).[82-83] In general, HILIC is used to separate hydrophilic glycopeptides from more hydrophobic peptides. However, this technique lacks specificity for glycopeptides; many non-glycosylated peptides may be separated along with glycopeptides, and as a result this method is not ideal for glycopeptide enrichment in complex biological samples. Several improvements have been made to HILIC enrichment methods, including using ion-pairing with trifluoroacetic acid (TFA) to increase the hydrophilicty of glycopeptides and enhance enrichment.[84-85]

1.2.3.2 Lectin enrichment

The most common method employed for glycoprotein/peptide separation uses lectins, which are glycan-binding proteins. Hundreds of lectins exist, and each lectin has inherent specificity for one or several glycans. As a result, lectin enrichment methods are not comprehensive, since there is no general lectin capable of binding to all glycans. While each lectin is highly specific for a particular carbohydrate moiety, the binding affinities of lectin-carbohydrate interactions are relatively low (low micromolar to millimolar range), which is another drawback of this method.[86-88] With low binding affinities, there are fewer options for washing away non-specifically bound peptides/proteins. Nevertheless, lectins have been used to successfully analyze glycoproteins in a variety of complex samples.[89-91] However, lectin-based methods on their own are not applicable for extracellular glycoprotein enrichment.

1.2.3.3 Boronic acid enrichment

Boronic acid chemistry has been applied for comprehensive glycoprotein enrichment. This technique is universal and can enrich all glycoproteins containing glycans with *cis*-diols.[92-93] Boronic acid forms reversible covalent interactions with *cis*-diols in glycans under basic conditions, and non-specifically bound glycoproteins can be washed away before eluting glycoproteins from boronic acid under acidic conditions.[94-95] This approach has been applied to enrich glycoproteins in serum and saliva.[96-97] Our group has implemented boronic acid conjugated magnetic beads for glycoprotein enrichment, which has proven to be highly effective in yeast.[98] Boronic acid enrichment serves as a promising avenue for comprehensive glycoprotein analysis; however these methods are unable to specifically enrich cell surface glycoproteins.

1.2.3.4 Hydrazide chemistry based enrichment

Hydrazide chemistry can be exploited for glycoprotein enrichment because it readily reacts with oxidized *cis*-diols in glycans. The first application of this utilized hydrazide attached to a solid support to capture oxidized glycoproteins from human serum, followed by proteolysis, removal of non-glycosylated peptides, and release using peptide N-glycosidase F (PNGase F).[99] A variation of this method performs the oxidation at the peptide level.[100] These methods have been applied to a variety of samples, including human plasma, islets, platelets, and saliva.[101-106] The oxidation conditions were then further optimized for compatibility with live-cell labeling: cell surface glycoproteins were labeled with biocytin hydrazide, digested and subsequently enriched with streptavidin beads.[107] This method has been widely applied for the specific identification of surface N-glycoproteins. However, the oxidation of carbohydrates is performed under relatively harsh conditions, which may act as an external stimulus for cells and cause side

reactions within proteins, which is the main drawback of this method.[108] Therefore, there is a need for an effective and specific method to enrich surface and extracellular glycoproteins that can be performed under mild conditions.

1.3 Goals of thesis

The overall goal of this thesis is to develop novel MS-based methods for global and accurate analysis of cell surface and extracellular glycoproteins. Glycosylation is one of the most common protein modifications in cells, and glycoproteins are critical for many cellular activities, especially extracellular interactions. Cell surface and extracellular glycoproteins are of particular interest because they frequently represent the cellular status, and serve as a promising source of biomarkers for disease progression and drug targets. Despite the prevalence and importance of protein glycosylation, it is extremely challenging to globally identify glycoproteins, due to the low abundance of many glycoproteins, heterogeneity of glycans, and complexity of biological samples. It is even more difficult to specifically characterize glycoproteins located only on the cell surface and in the secretome. The work described here provides novel MS-based methods to comprehensively analyze extracellular glycoproteins. Effective methods to investigate surface and secretory glycoproteins on a large scale enable a better understanding of glycoprotein function and the molecular mechanisms of disease. The methods presented in this dissertation provide the foundation for the future discovery of specific glycoproteins as drug targets and biomarkers for disease.

1.4 References

1. Milo, R., What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **2013,** *35* (12), 1050-1055.

2. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass-spectrometry of large biomolecules. *Science* **1989,** *246* (4926), 64-71.

3. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y., Protein and polymer analysis up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** *2*, 151-153.

4. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H., Characterization of peptides bound to the class-I MHC molecule HLA-A2.1 by mass-spectrometry. *Science* **1992,** *255* (5049), 1261-1263.

5. Holliman, C. L.; Rempel, D. L.; Gross, M. L., Detection of high mass-to-charge ions by Fourier-transform mass-spectrometry. *Mass Spectrom. Rev.* **1994,** *13* (2), 105-132.

6. Loo, J. A., Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **1997,** *16* (1), 1-23.

7. Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J., Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **2013,** *10* (4), 332-334.

8. Skinner, O. S.; Catherman, A. D.; Early, B. P.; Thomas, P. M.; Compton, P. D.; Kelleher, N. L., Fragmentation of integral membrane proteins in the gas phase. *Anal. Chem.* **2014,** *86* (9), 4627-4634.

9. Zhou, M. W.; Wysocki, V. H., Surface induced dissociation: Dissecting noncovalent protein complexes in the gas phase. *Accounts Chem. Res.* **2014,** *47* (4), 1010-1018.

10. Yates, J. R., The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J. Am. Chem. Soc.* **2013,** *135* (5), 1629-1640.

11. Neilson, K. A.; Scafaro, A. P.; Chick, J. M.; George, I. S.; Van Sluyter, S. C.; Gygi, S. P.; Atwell, B. J.; Haynes, P. A., The influence of signals from chilled roots on the proteome of shoot tissues in rice seedlings. *Proteomics* **2013,** *13* (12-13), 1922-1933.

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

13. Yao, X. D.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C., Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus. *Anal. Chem.* **2001,** *73* (13), 2836-2842.

14. Zhu, Z. K.; Go, E. P.; Desaire, H., Absolute quantitation of glycosylation site occupancy using isotopically labeled standards and LC-MS. *J. Am. Soc. Mass Spectrom.* **2014,** *25* (6), 1012-1017.

15. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-355.

16. Marguerat, S.; Schmidt, A.; Codlin, S.; Chen, W.; Aebersold, R.; Bahler, J., Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **2012,** *151* (3), 671-683.

17. Kelleher, N. L., Top-down proteomics. *Anal. Chem.* **2004,** *76* (11), 196a-203a.

18. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M. X.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-U141.

19. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W., Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *J Am Chem Soc* **1999,** *121* (4), 806-812.

20. Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L., Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: Human histone H4. *Anal. Chem.* **2006,** *78* (13), 4271-4280.

21. Du, Y.; Parks, B. A.; Sohn, S.; Kwast, K. E.; Kelleher, N. L., Top-down approaches for measuring expression ratios of intact yeast proteins using Fourier transform mass spectrometry. *Anal. Chem.* **2006,** *78* (3), 686-694.

22. Waanders, L. F.; Hanke, S.; Mann, M., Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of the American Society for Mass Spectrometry* **2007,** *18* (11), 2058-2064.

23. Yates, J. R.; Ruse, C. I.; Nakorchevsky, A., Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* **2009,** *11*, 49-79.

24. Forbes, A. J.; Mazur, M. T.; Patel, H. M.; Walsh, C. T.; Kelleher, N. L., Toward efficient analysis of > 70 kDa proteins with 100% sequence coverage. *Proteomics* **2001,** *1* (8), 927-933.

25. Wu, C.; Tran, J. C.; Zamdborg, L.; Durbin, K. R.; Li, M. X.; Ahlf, D. R.; Early, B. P.; Thomas, P. M.; Sweedler, J. V.; Kelleher, N. L., A protease for 'middle-down' proteomics. *Nat. Methods* **2012,** *9* (8), 822-+.

26. Tsiatsiani, L.; Heck, A. J. R., Proteomics beyond trypsin. *Febs J* **2015,** *282* (14), 2612-2626.

27. Kalli, A.; Hakansson, K., Electron capture dissociation of highly charged proteolytic peptides from Lys N, Lys C and Glu C digestion. *Mol. Biosyst.* **2010,** *6* (9), 1668-1681.

28. Sidoli, S.; Lin, S.; Karch, K. R.; Garcia, B. A., Bottom-up and middle-down proteomics have comparable accuracies in defining histone post-translational modification relative abundance and stoichiometry. *Anal. Chem.* **2015,** *87* (6), 3129-3133.

29. Kalli, A.; Sweredoski, M. J.; Hess, S., Data-dependent middle-down nano-liquid chromatography-electron capture dissociation-tandem mass spectrometry: an application for the analysis of unfractionated histones. *Anal. Chem.* **2013,** *85* (7), 3501-3507.

30. Cannon, J.; Lohnes, K.; Wynne, C.; Wang, Y.; Edwards, N.; Fenselau, C., High-throughput middle-down analysis using an orbitrap. *J Proteome Res* **2010,** *9* (8), 3886-3890.

31. Xu, P.; Peng, J. M., Characterization of polyubiquitin chain structure by middle-down mass spectrometry. *Anal. Chem.* **2008,** *80* (9), 3438-3444.

32. Wolters, D. A.; Washburn, M. P.; Yates, J. R., An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001,** *73* (23), 5683-5690.

33. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

34. Yates, J. R.; Eng, J. K.; Mccormack, A. L.; Schieltz, D., Method to correlate tandem mass-spectra of modified peptides to amino-acid-sequences in the protein database. *Anal. Chem.* **1995,** *67* (8), 1426-1436.

35. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** *20* (18), 3551-3567.

36. Steen, H.; Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **2004,** *5* (9), 699-711.

37. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999,** *17* (7), 676-682.

38. Washburn, M. P.; Wolters, D.; Yates, J. R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** *19* (3), 242-247.

39. Mann, M., Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **2006,** *7* (12), 952-958.

40. Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T., Accurate quantitation of protein expression and site-specific phosphorylation. *P Natl Acad Sci USA* **1999,** *96* (12), 6591-6596.

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

42. Ross, P. L.; Huang, Y. L. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** *3* (12), 1154-1169.

43. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C., Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003,** *75* (8), 1895-1904.

44. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999,** *17* (10), 994-999.

45. Shiio, Y.; Aebersold, R., Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nat Protoc* **2006,** *1* (1), 139-145.

46. Wang, Y. C.; Jobe, S. M.; Ding, X. K.; Choo, H. J.; Archer, D. R.; Mi, R. J.; Ju, T. Z.; Cummings, R. D., Platelet biogenesis and functions require correct protein O-glycosylation. *Proc. Natl. Acad. Sci. U. S. A.* **2012,** *109* (40), 16143-16148.

47. Dai, L. Z.; Peng, C.; Montellier, E.; Lu, Z. K.; Chen, Y.; Ishii, H.; Debernardi, A.; Buchou, T.; Rousseaux, S.; Jin, F. L.; Sabari, B. R.; Deng, Z. Y.; Allis, C. D.; Ren, B.; Khochbin, S.; Zhao, Y. M., Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat. Chem. Biol.* **2014,** *10* (5), 365-370.

48. Ma, J. F.; Hart, G. W., Protein O-GlcNAcylation in diabetes and diabetic complications. *Expert Rev. Proteomics* **2013,** *10* (4), 365-380.

49. Xiao, Y.; Lee, T.; Latham, M. P.; Warner, L. R.; Tanimoto, A.; Pardi, A.; Ahn, N. G., Phosphorylation releases constraints to domain motion in ERK2. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (7), 2506-2511.

50. Bullen, J. W.; Balsbaugh, J. L.; Chanda, D.; Shabanowitz, J.; Hunt, D. F.; Neumann, D.; Hart, G. W., Cross-talk between two essential nutrient-sensitive enzymes O-GlcNAc transferase (OGT) and AMP-activated protein kinase (AMPK). *J. Biol. Chem.* **2014,** *289* (15), 10592-10606.

51. Nie, S.; Lo, A.; Wu, J.; Zhu, J. H.; Tan, Z. J.; Simeone, D. M.; Anderson, M. A.; Shedden, K. A.; Ruffin, M. T.; Lubman, D. M., Glycoprotein biomarker panel for pancreatic cancer discovered by quantitative proteomics analysis. *J. Proteome Res.* **2014,** *13* (4), 1873-1884.

52. Wu, J.; Xie, X. L.; Nie, S.; Buckanovich, R. J.; Lubman, D. M., Altered expression of sialylated glycoproteins in ovarian cancer sera using lectin-based ELISA assay and quantitative glycoproteomics analysis. *J. Proteome Res.* **2013,** *12* (7), 3342-3352.

53. Gilgunn, S.; Conroy, P. J.; Saldova, R.; Rudd, P. M.; O'Kennedy, R. J., Aberrant PSA glycosylation - A sweet predictor of prostate cancer. *Nat. Rev. Urol.* **2013,** *10* (2), 99-107.

54. Oda, Y.; Nagasu, T.; Chait, B. T., Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **2001,** *19* (4), 379-382.

55. Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M., Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **2002,** *20* (3), 301-305.

56. Yang, W. M.; Shah, P.; Eshghi, S. T.; Yang, S.; Sun, S. S.; Ao, M. H.; Rubin, A.; Jackson, J. B.; Zhang, H., Glycoform analysis of recombinant and human immunodeficiency virus envelope protein GP120 *via* higher energy collisional dissociation and spectral-aligning strategy. *Anal. Chem.* **2014,** *86* (14), 6959-6967.

57. Chen, W. X.; Smeekens, J. M.; Wu, R. H., Comprehensive analysis of protein N-glycosylation sites by combining chemical deglycosylation with LC-MS. *J Proteome Res* **2014,** *13* (3), 1466-1473.

58. Wu, R.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat Meth* **2011,** *8* (8), 677-683.

59. Mann, M.; Jensen, O. N., Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003,** *21* (3), 255-261.

60. Witze, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G., Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **2007,** *4* (10), 798-806.

61. Zhao, Y. Y.; Takahashi, M.; Gu, J. G.; Miyoshi, E.; Matsumoto, A.; Kitazume, S.; Taniguchi, N., Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci.* **2008,** *99* (7), 1304-1310.

62. Ohtsubo, K.; Marth, J. D., Glycosylation in cellular mechanisms of health and disease. *Cell* **2006,** *126* (5), 855-867.

63. Apweiler, R.; Hermjakob, H.; Sharon, N., On the frequency of protein glycosylation, as deduced from analysis of the Swiss-Prot database. *Biochim. Biophys. Acta-Gen. Subj.* **1999,** *1473* (1), 4-8.

64. Bause, E., Structural requirements of n-glycosylation of proteins - studies with proline peptides as conformational probes. *Biochemical Journal* **1983,** *209* (2), 331-336.

65. Rudd, P. M.; Elliott, T.; Cresswell, P.; Wilson, I. A.; Dwek, R. A., Glycosylation and the immune system. *Science* **2001,** *291* (5512), 2370-2376.

66. Varki, A., *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press: 2009.

67. Vigerust, D. J., Protein glycosylation in infectious disease pathobiology and treatment. *Cent. Eur. J. Biol.* **2011,** *6* (5), 802-816.

68. Dube, D. H.; Bertozzi, C. R., Glycans in cancer and inflammation. Potential for therapeutics and diagnostics. *Nat Rev Drug Discov* **2005,** *4* (6), 477-488.

69. Woods, R. J.; Edge, C. J.; Dwek, R. A., Protein surface oligosaccharides and protein function. *Nat Struct Biol* **1994,** *1* (8), 499-501.

70. Haltiwanger, R. S.; Lowe, J. B., Role of glycosylation in development. *Annu Rev Biochem* **2004,** *73*, 491-537.

71. Grimm, D.; Bauer, J.; Pietsch, J.; Infanger, M.; Eucker, J.; Eilles, C.; Schoenberger, J., Diagnostic and therapeutic use of membrane proteins in cancer cells. *Curr Med Chem* **2011,** *18* (2), 176-190.

72. Rucevic, M.; Hixson, D.; Josic, D., Mammalian plasma membrane proteins as potential biomarkers and drug targets. *Electrophoresis* **2011,** *32* (13), 1549-1564.

73. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M., Drug-target network. *Nat Biotechnol* **2007,** *25* (10), 1119-1126.

74. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat Rev Drug Discov* **2002,** *1* (9), 727-730.

75. Stastna, M.; Van Eyk, J. E., Secreted proteins as a fundamental source for biomarker discovery. *Proteomics* **2012,** *12* (4-5), 722-735.

76. Sokoll, L. J.; Chan, D. W., Prostate-specific antigen - Its discovery and biochemical characteristics. *Urol Clin N Am* **1997,** *24* (2), 253-&.

77. Kim, E. H.; Misek, D. E., Glycoproteomics-based identification of cancer biomarkers. *International Journal of Proteomics* **2011,** *2011*, 10.

78. Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R., The quantitative proteome of a human cell line. *Mol Syst Biol* **2011,** *7*.

79. Picotti, P.; Aebersold, R.; Domon, B., The implications of proteolytic background for shotgun proteomics. *Mol Cell Proteomics* **2007,** *6* (9), 1589-1598.

80. Stavenhagen, K.; Hinneburg, H.; Thaysen-Andersen, M.; Hartmann, L.; Silva, D. V.; Fuchser, J.; Kaspar, S.; Rapp, E.; Seeberger, P. H.; Kolarich, D., Quantitative mapping of glycoprotein micro-heterogeneity and macro-heterogeneity: an evaluation of mass spectrometry signal strengths using synthetic peptides and glycopeptides. *J Mass Spectrom* **2013,** *48* (6), 627-639.

81. Spiro, R. G., Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **2002,** *12* (4), 43R-56R.

82. Hagglund, P.; Bunkenborg, J.; Elortza, F.; Jensen, O. N.; Roepstorff, P., A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *J Proteome Res* **2004,** *3* (3), 556-566.

83. Zauner, G.; Deelder, A. M.; Wuhrer, M., Recent advances in hydrophilic interaction liquid chromatography (HILIC) for structural glycomics. *Electrophoresis* **2011,** *32* (24), 3456-3466.

84. Mysling, S.; Palmisano, G.; Hojrup, P.; Thaysen-Andersen, M., Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. *Anal. Chem.* **2010,** *82* (13), 5598-5609.

85. Jensen, P. H.; Mysling, S.; Højrup, P.; Jensen, O. N., Glycopeptide enrichment for MALDI-TOF mass spectrometry analysis by hydrophilic interaction liquid chromatography solid phase extraction (HILIC SPE). In *Mass Spectrometry of Glycoproteins: Methods and Protocols*, Kohler, J. J.; Patrie, S. M., Eds. Humana Press: Totowa, NJ, 2013; pp 131-144.

86. Gupta, D.; Dam, T. K.; Oscarson, S.; Brewer, C. F., Thermodynamics of lectin-carbohydrate interactions. *J Biol Chem* **1997,** *272* (10), 6388-6392.

87. Mandal, D. K.; Kishore, N.; Brewer, C. F., Thermodynamics of lectin-carbohydrate interactions - titration microcalorimetry measurements of the binding of n-linked carbohydrates and ovalbumin to Concanavalin-A. *Biochemistry-Us* **1994,** *33* (5), 1149-1156.

88. Dam, T. K.; Roy, R.; Das, S. K.; Oscarson, S.; Brewer, C. F., Binding of multivalent carbohydrates to Concanavalin A and *Dioclea grandiflora* lectin - Thermodynamic analysis of the "multivalency effect". *J Biol Chem* **2000,** *275* (19), 14223-14230.

89. Kaji, H.; Saito, H.; Yamauchi, Y.; Shinkawa, T.; Taoka, M.; Hirabayashi, J.; Kasai, K.; Takahashi, N.; Isobe, T., Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* **2003,** *21* (6), 667-672.

90. Madera, M.; Mechref, Y.; Novotny, M. V., Combining lectin microcolumns with high-resolution separation techniques for enrichment of glycoproteins and glycopeptides. *Anal. Chem.* **2005,** *77* (13), 4081-4090.

91. Drake, R. R.; Schwegler, E. E.; Malik, G.; Diaz, J.; Block, T.; Mehta, A.; Semmes, O. J., Lectin capture strategies combined with mass spectrometry for the discovery of serum glycoprotein biomarkers. *Mol Cell Proteomics* **2006,** *5* (10), 1957-1967.

92. Siegel, D., Applications of reversible covalent chemistry in analytical sample preparation. *Analyst* **2012,** *137* (23), 5457-5482.

93. Bull, S. D.; Davidson, M. G.; Van den Elsen, J. M. H.; Fossey, J. S.; Jenkins, A. T. A.; Jiang, Y. B.; Kubo, Y.; Marken, F.; Sakurai, K.; Zhao, J. Z.; James, T. D., Exploiting the reversible covalent bonding of boronic acids: recognition, sensing, and assembly. *Accounts of Chemical Research* **2013,** *46* (2), 312-326.

94. Sparbier, K.; Koch, S.; Kessler, I.; Wenzel, T.; Kostrzewa, M., Selective isolation of glycoproteins and glycopeptides for MALDI-TOF MS detection supported by magnetic particles. *Journal of Biomolecular Techniques : JBT* **2005,** *16* (4), 407-413.

95. Xu, Y. W.; Wu, Z. X.; Zhang, L. J.; Lu, H. J.; Yang, P. Y.; Webley, P. A.; Zhao, D. Y., Highly specific enrichment of glycopeptides using boronic acid-functionalized mesoporous silica. *Anal. Chem.* **2009,** *81* (1), 503-508.

96. Sparbier, K.; Wenzel, T.; Kostrzewa, M., Exploring the binding profiles of ConA, boronic acid and WGA by MALDI-TOF/TOF MS and magnetic particles. *J Chromatogr B* **2006,** *840* (1), 29-36.

97. Xu, Y.; Bailey, U. M.; Punyadeera, C.; Schulz, B. L., Identification of salivary N-glycoproteins and measurement of glycosylation site occupancy by boronate glycoprotein enrichment and liquid chromatography/electrospray ionization tandem mass spectrometry. *Rapid Commun Mass Sp* **2014,** *28* (5), 471-482.

98. Chen, W. X.; Smeekens, J. M.; Wu, R. H., A universal chemical enrichment method for mapping the yeast n-glycoproteome by mass spectrometry (MS). *Mol Cell Proteomics* **2014,** *13* (6), 1563-1572.

99. Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R., Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* **2003,** *21* (6), 660-666.

100. Tian, Y. A.; Zhou, Y.; Elliott, S.; Aebersold, R.; Zhang, H., Solid-phase extraction of N-linked glycopeptides. *Nat Protoc* **2007,** *2* (2), 334-339.

101. Liu, T.; Qian, W. J.; Gritsenko, M. A.; Camp, D. G.; Monroe, M. E.; Moore, R. J.; Smith, R. D., Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J Proteome Res* **2005,** *4* (6), 2070-2080.

102. Liu, Y. S.; Huttenhain, R.; Surinova, S.; Gillet, L. C. J.; Mouritsen, J.; Brunner, R.; Navarro, P.; Aebersold, R., Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics* **2013,** *13* (8), 1247-1256.

103. Zhou, Y.; Aebersold, R.; Zhang, H., Isolation of N-linked glycopeptides from plasma. *Anal. Chem.* **2007,** *79* (15), 5826-5837.

104. Danzer, C.; Eckhardt, K.; Schmidt, A.; Fankhauser, N.; Ribrioux, S.; Wollscheid, B.; Muller, L.; Schiess, R.; Zullig, R.; Lehmann, R.; Spinas, G.; Aebersold, R.; Krek, W., Comprehensive Description of the N-Glycoproteome of Mouse Pancreatic beta-Cells and Human Islets. *J Proteome Res* **2012,** *11* (3), 1598-1608.

105. Lewandrowski, U.; Moebius, J.; Walter, U.; Sickmann, A., Elucidation of N-glycosylation sites on human platelet proteins - A glycoproteomic approach. *Mol Cell Proteomics* **2006,** *5* (2), 226-233.

106. Ramachandran, P.; Boontheung, P.; Xie, Y. M.; Sondej, M.; Wong, D. T.; Loo, J. A., Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. *J Proteome Res* **2006,** *5* (6), 1493-1503.

107. Wollscheid, B.; Bausch-Fluck, D.; Henderson, C.; O'Brien, R.; Bibel, M.; Schiess, R.; Aebersold, R.; Watts, J. D., Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* **2009,** *27* (4), 378-386.

108. Huang, J. F.; Qin, H. Q.; Sun, Z.; Huang, G.; Mao, J. W.; Cheng, K.; Zhang, Z.; Wan, H.; Yao, Y. T.; Dong, J.; Zhu, J.; Wang, F. J.; Ye, M. L.; Zou, H. F., A peptide N-

terminal protection strategy for comprehensive glycoproteome analysis using hydrazide chemistry based method. *Sci Rep-Uk* **2015,** *5*.

CHAPTER 2. MASS SPECTROMETRIC ANALYSIS OF THE CELL SURFACE N-GLYCOPROTEOME

Adapted with permission from Springer

Smeekens, J.M., Chen, W., Wu, R. Mass spectrometric analysis of the cell surface N-glycoproteome by combining metabolic labeling and click chemistry. *Journal of the American Society for Mass Spectrometry*. 2015, 26 (4), 604-614.

2.1 Introduction

Glycosylation is one of the most common and critical protein modifications in cells; it has been estimated that about 50% of proteins expressed in a cell are glycosylated.[1] Additionally, nearly all extracellular proteins are glycosylated, including cell surface and membrane proteins. These extracellular glycoproteins play crucial roles in cell-cell interactions and immune response.[2-3] Abnormal glycosylation is correlated with the development of disease, including cancer and Alzheimers.[4-5] Extracellular glycoproteins often represent the diseased or developmental status of the cell, and therefore are a promising source of non-invasive biomarkers.[6] Furthermore, cell surface glycoproteins are commonly used as therapeutic targets; the majority of FDA approved drugs target surface proteins.[7-8] Therefore, there is great interest in comprehensively analyzing cell surface glycoproteins, but there is a lack of effective methods. As described in Chapter 1.2.1, extracellular glycoprotein analysis is extremely challenging, and requires effective enrichment from non-modified and intracellular proteins.

Several years ago a very elegant cell surface capturing method was reported for cell surface glycoprotein analysis.[9] Glycoproteins on living cells were oxidized, bound to biotin through a bifunctional linker molecule, biocytin hydrazide, and enriched by avidin for MS analysis. However, the oxidation reaction conditions are not under physiological conditions and therefore can act as an external stimulus to cells. Effective methods will profoundly advance the analysis of cell surface glycoproteins and provide insight into glycoprotein function.

In this work we have developed an effective MS-based method to identify cell surface glycoproteins comprehensively and site-specifically. A sugar analog containing a biologically inert but chemically functional azido group was fed to cells to label cell surface glycoproteins according to a previous method.[10] Surface glycoproteins containing the functional group were subsequently bound to biotin through copper-free click chemistry under mild physiological conditions. Further separation and enrichment by exploiting the strong interaction between biotin and avidin allowed the global analysis of cell surface glycoproteins.

2.2 Experimental Methods

2.2.1 Cell culture and surface glycoprotein labeling

HEK 293T cells were grown in Dulbecco's Modified Eagle Medium (DMEM) containing 10 % fetal bovine serum (FBS). Once cells reached 10 % confluency, medium was changed to DMEM containing 10 % FBS and 50 μM N-azidoacetylgalactosamine (GalNAz). Cells were incubated for 3 days until confluency was ~80%. After metabolic labeling, cells were washed with phosphate buffered saline (PBS) two times and 100 μM

dibenzocyclooctyne (DBCO)-sulfo-biotin in PBS was added into the culture flasks. Cells were incubated for one hour with gentle agitation at 4 ºC and then harvested by scraping in PBS and centrifugation at 300 g for 5 minutes. The supernatant was discarded and the cell pellet containing about $4x10^7$ cells was washed twice with PBS containing 10 mM dithiothreitol (DTT).

2.2.2 Cell lysis and membrane protein extraction

The cell pellet was incubated in a buffer containing 150 mM NaCl, 50 mM 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (HEPES) (pH=7.4), 25 ug/mL digitonin, and Roche protease inhibitor (1 tablet per 10 mL) with end-over-end rotation for 10 minutes at 4°C. After incubation, samples were centrifuged at 2,000 g for 10 minutes. Cell pellets were washed with the buffer twice and subsequently lysed with the MiniBeadbeater (Biospec) in a buffer containing 10 mM HEPES, 1.5 mM $MgCl_2$ and 10 mM KCl. The resulting solutions were centrifuged at 2,500 g for 10 minutes, and the supernatant was collected and centrifuged at 16,000 g for 30 minutes. The membrane rich pellet was collected and washed two times with the lysis buffer. The pellet was further incubated in 0.1 M sodium carbonate solution containing 1 mM ethylenediaminetetraacetic acid (EDTA) on ice for 30 minutes, followed by centrifugation at 16,000 g for 15 minutes. Sodium carbonate incubations and subsequent centrifugations were repeated once. The membrane rich pellets were incubated with shaking in a buffer containing 4 M urea, 100 mM NaCl, 10 mM HEPES, and 1 mM EDTA for 30 minutes at room temperature, and then samples were centrifuged at 16,000 g for 15 minutes. The urea buffer wash was repeated once. Solubilization buffer

containing 100 mM PBS and 1 % NP-40 was added to cell pellets and incubated end-over-end overnight at room temperature.

2.2.3 Glycopeptide preparation and enrichment

Solubilized samples were centrifuged at 16,000 g for 15 minutes and supernatants were collected. Disulfide bonds within proteins were reduced with 5 mM DTT (56 °C, 25 minutes) and subsequently alkylated with 14 mM iodoacetamide (room temperature, 30 minutes in the dark). After reduction and alkylation, proteins were purified with the methanol chloroform protein precipitation method [11]. Four volumes of methanol, one volume of chloroform and three volumes of water were added to one volume of the protein sample, and the mixture was vortexed. The sample was centrifuged at 5,000 g for 20 minutes. The proteins remained at the phase boundary between the methanol and chloroform layers. The methanol layer above the sample was removed. Four volumes of methanol were added, and the mixture was vortexed again. The sample was centrifuged again at 5,000 g for 15 minutes. The supernatant was removed without disturbing the pellet, and the pellet was dried.

The resulting ~2 mg protein samples were digested overnight at a protein:trypsin ratio of ~100:1 in 2 mL of buffer containing 50 mM HEPES (pH=8.5), 0.1 M urea, and 5% ACN. The next day, the digestion was quenched by the addition of trifluoroacetic acid (TFA) to a final concentration of 0.1% and digested peptides were purified with a 200 mg Sep-Pak tC18 cartridge. Purified samples were dried and enriched with 200 µL of NeutrAvidin bead slurry end-over-end for 30 minutes at 37 °C. After enrichment, the peptide sample was transferred to a spin column and beads were washed ten times with

400 µL PBS and once with water. Peptides were eluted from the beads twice by two minute incubations with 200 µL of 8 M guanidine (pH=1.5) at 56 °C. Combined eluates were purified on a 50 mg Sep-Pak tC18 cartridge. Enriched peptides were dried thoroughly before enzymatic deglycosylation with 8 units of peptide-N-glycosidase F (PNGase F, Sigma-Aldrich) in 40 µL of buffer containing 40 mM NH_4HCO_3 in heavy-oxygen water ($H_2{}^{18}O$) for three hours at 37 °C. The deglycosylation reaction was quenched with formic acid (FA) and purified with the stage tip method. Proteins were eluted into three samples using 20%, 50% and 80% ACN each containing 1% HOAc.

2.2.4 LC-MS/MS analysis

Purified samples were dried and resuspended in a solvent containing 5% ACN and 4% FA, and 4 µL were loaded onto a C18-packed microcapillary column (Magic C18AQ, 5 µm, 200 Å, 100 µm x 16 cm) using a WPS-3000TPL RS autosampler (Thermostatted Pulled Loop Rapid Separation Nano/Capillary Autosampler, Dionex). Peptides were separated by reversed-phase chromatography using an UltiMate 3000 binary pump with a 110 minute gradient that varied for three samples. The first sample had a gradient of 4-25% ACN (0.125% FA), the second sample's gradient was 10-38% ACN (0.125% FA) and the third sample had a gradient of 15-50% ACN (0.125% FA). Samples were detected in a hybrid dual-cell quadrupole linear ion trap – Orbitrap mass spectrometer (LTQ Orbitrap Elite, ThermoFisher) using a data-dependent Top 20 method. Each cycle included one full MS scan (resolution: 60,000) in the Orbitrap at the Automatic Gain Control (AGC) target of 1 million, followed by up to 20 MS/MS of the most intense ions in the LTQ. Selected ions were excluded from being further sequenced

for 90 seconds. Ions with a single or unassigned charge were not fragmented. Maximum ion accumulation times were 1000 ms for each full MS scan and 50 ms for MS/MS scans.

2.2.5 Database searches and filtering

The raw files recorded by MS were converted into mzXML format. Precursors for MS/MS fragmentation were checked for incorrect monoisotopic peak assignments while refining precursor ion mass measurements. The SEQUEST algorithm [12] (version 28) was used to search and match all MS/MS spectra against a database encompassing sequences of all proteins in the Uniprot Human (*Homo sapiens*) Database containing common contaminants such as keratins. Each protein sequence was listed in both forward and reversed orientations to estimate the false discovery rate (FDR) of glycopeptide identification.[13-14] A 20 ppm precursor mass tolerance and 1.0 Da product ion mass tolerance were used in the database search. Other selected parameters were fully tryptic digestion, up to two missed cleavages, variable modifications including oxidation of methionine (+15.9949) and ^{18}O tag of Asn (+2.9883), and fixed modifications including carbamidomethylation of cysteine (+57.0214).

In order to evaluate and further control FDRs of glycopeptide identification, the target-decoy method was employed.[13-14] Linear discriminant analysis (LDA) was utilized to distinguish correct and incorrect peptide identifications using numerous parameters including XCorr, ΔCn, and precursor mass error.[15] Separate linear discriminant models were trained for each raw file using forward and reversed peptide sequences to provide positive and negative training data. This approach is similar to other methods in the literature.[16-17] After scoring, peptides less than six amino acids in length were discarded

and peptide spectral matches were filtered to a less than 1% FDR based on the number of decoy sequences in the final data set. The dataset was restricted to glycopeptides when determining FDRs.

2.2.6 Glycosylation site localization

To localize glycosylation sites and obtain a level of confidence corresponding to the identification, we applied a probabilistic algorithm that considers all potential glycosylation sites on a peptide and uses the presence or absence of experimental fragment ions unique to each to obtain a ModScore.[18] The ModScore, which is similar to Ascore, indicates the likelihood that the best site match is correct when compared with the next best match.[18] If only one glycosylation site is possible, a value of 1000 is assigned to the site. We considered sites with a score $\geq$19 ($P \leq 0.01$) to be confidently localized.

2.3 Results and Discussion

2.3.1 Glycopeptide labeling and enrichment

Sugar analogs have frequently been used to discover glycotransferase inhibitors.[19] Some sugar analogs can be used by glycosyltransferases to modify glycans within mammalian proteins. By taking advantage of this discovery, scientists have extensively investigated culturing cells with sugar analogs bearing various functional groups, including azido, alkyl and aldehyde groups.[10, 20-22] The azido functional group in the sugar analog serves as a chemical handle for the click chemistry reaction which will provide further insight into protein function and cellular activities.

In recent years, metabolic labeling has been employed for *in vivo* imaging experiments and valuable information has been obtained regarding the location of the sugar analogs and their relative abundance changes based on the fluorescence signal and corresponding intensity changes.[21, 23] In previous studies, GalNAz has been used in mammalian cells to modify proteins, resulting in azido-containing glycoproteins.[24] In this experiment, GalNAz was added into DMEM and incubated with HEK 293T cells in order to metabolically label glycoproteins.

Cell surface glycoproteins with the functional azido group were selectively bound to biotin through copper-free click chemistry, as shown in Figure 2.1. The copper-free click chemistry reaction between DBCO and the functional azido group is very specific, rapid and efficient.[25-26] More importantly, the reaction can occur under physiological conditions, without toxic heavy metal ions (Cu(I) and Cu(II)) that are frequently used as a catalyst in the traditional copper-catalyzed azide-alkyne cycloaddition (CuAAC), which allows surface glycoproteins on living cells to be tagged and minimizes external stimuli.[27] After metabolic labeling and surface glycoprotein tagging, cells were lysed and proteins were extracted and digested. Based on the strong and specific interaction between biotin and avidin, glycopeptides tagged with biotin were selectively enriched by incubation with NeutrAvidin conjugated agarose beads. After incubation, the beads were washed to remove non-biotinylated peptides, and enriched glycopeptides were subsequently eluted from the beads.

Enriched glycopeptides were dried for at least 24 hours in a vacuum concentrator and then treated with PNGase F in heavy-oxygen water. This enzyme cleaves N-glycans and converts asparagine (Asn) to aspartic acid (Asp) in the process. The deamination of

Asn can also occur *in vivo* and during sample preparation. By performing this enzymatic deglycosylation in heavy-oxygen water, newly formed Asp is labeled with ^{18}O, which helps distinguish *bona fide* glycosylation sites from those caused by non-enzymatic deamination.[28-29] As a result, N-glycopeptides and their corresponding glycosylation sites were confidently identified and localized.

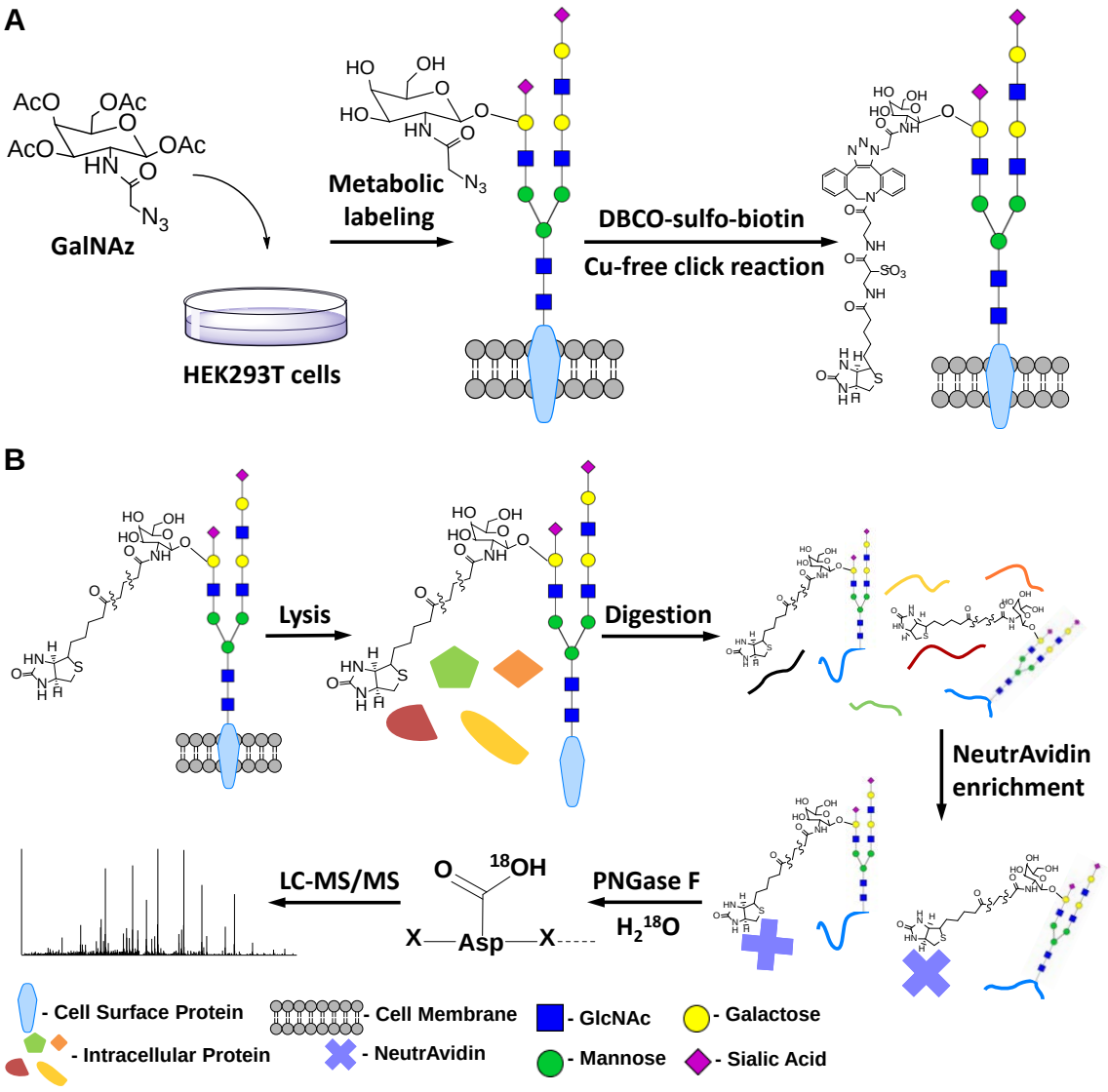

Figure 2.1. Principle of the cell surface glycoprotein enrichment method including (A) the metabolic labeling and click chemistry reaction, and (B) glycopeptide separation and analysis.

2.3.2 Glycopeptide identification and site localization

Using this novel strategy, 144 unique N-glycosylated peptides were identified. Figure 2.2 shows example tandem mass spectra of three identified N-glycopeptides, EN#TSDPSLVIAFGR, GHTLTLN#FTR, and YSVQLMSFVYN#LSDTHLFPN#ASSK (# denotes the glycosylation site) from lysosome-associated membrane glycoprotein 1 (LAMP1), which is a single-pass type I membrane protein. It presents carbohydrate ligands to selectins and is involved in tumor cell metastasis.[30] These peptides were confidently identified with XCorr values of 4.11, 2.45, and 4.50, respectively. The high mass accuracy of each identification is represented by their ppm values of 0.84, 0.14 and -0.16. The full sequence of this protein is shown in Figure 2.2D, with identified peptides highlighted. Corresponding glycosylation sites were confidently localized at N84, N103, N121 and N130, with ModScore values of 1000 for each. Additionally, all sites contain the consensus motif Asn-X-Ser/Thr, where X is any amino acid residue except proline. All glycosylation sites identified in this work are listed in a supplementary table online at doi.org/10.1007/s13361-014-1016-7.

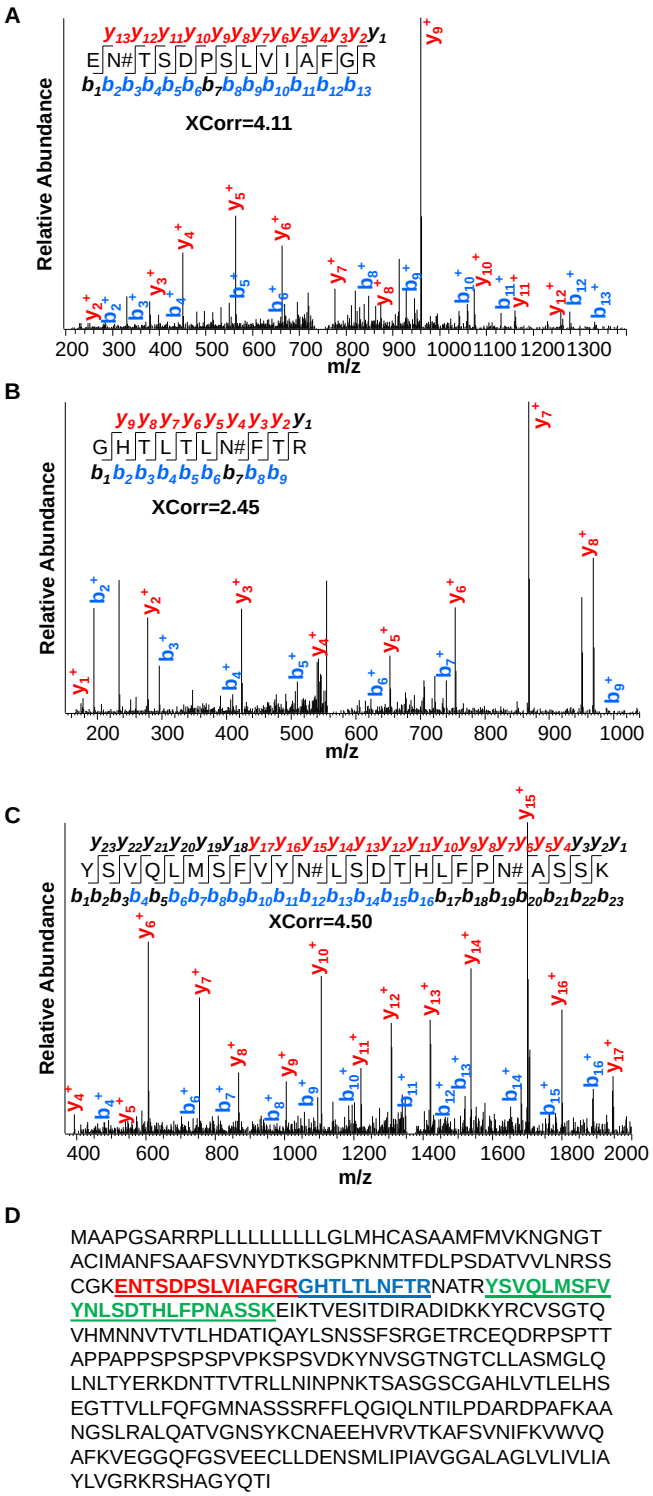

Figure 2.2. Tandem mass spectra of three peptides from the LAMP1 protein, including (A) EN#TSDPSLVIAFGR, (B) GHTLTLN#FTR, and (C) YSVQLMSFVYN#LSDTHLFPN#ASSK (# denotes the glycosylation site). The complete protein sequence with the highlighted identified peptides is shown in (D).

The high resolution and mass accuracy of the Orbitrap mass spectrometer allowed for the very confident N-glycopeptide identifications. The mass accuracy distribution of all identified N-glycopeptides is displayed in Figure 2.3A. Clearly, the vast majority of identified glycopeptides have a mass accuracy of less than 3 ppm. The mass accuracy results show the high confidence level associated with glycopeptide identifications. This is further confirmed by the XCorr values, which are shown in Figure 2.3B. Most of the XCorr values are greater than 3, and only several of them are less than 2. We manually checked glycopeptide identifications with relatively low XCorr values, and all of them were relatively short peptides whose fragments matched very well with corresponding theoretical spectra. The low XCorr values assigned to short peptides are due to the fact that the peptide matches are normally biased for longer peptides, *i.e.,* higher XCorr values for longer peptides, since there are more potential fragments and therefore more possible matches.[31]

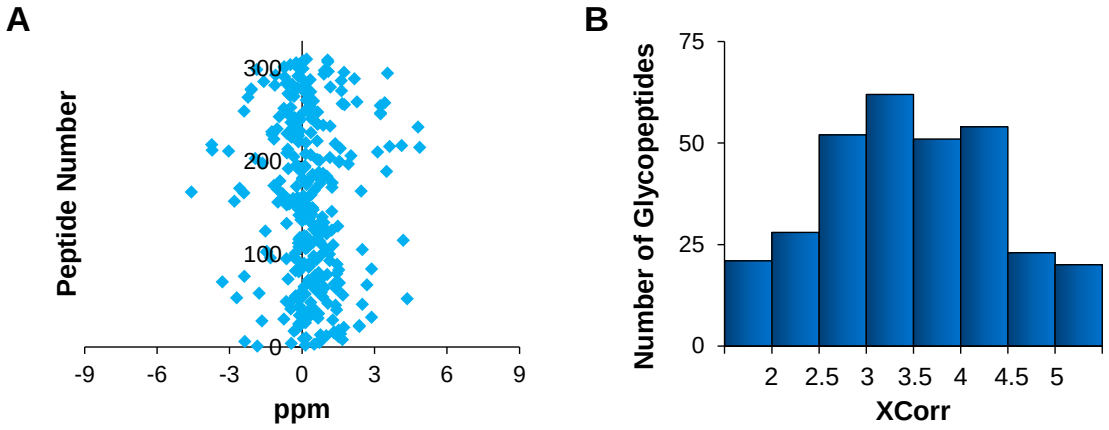

Figure 2.3. Distribution of (A) ppm and (B) XCorr values assigned for each glycopeptide identification. The bin size for XCorr analysis is 0.5.

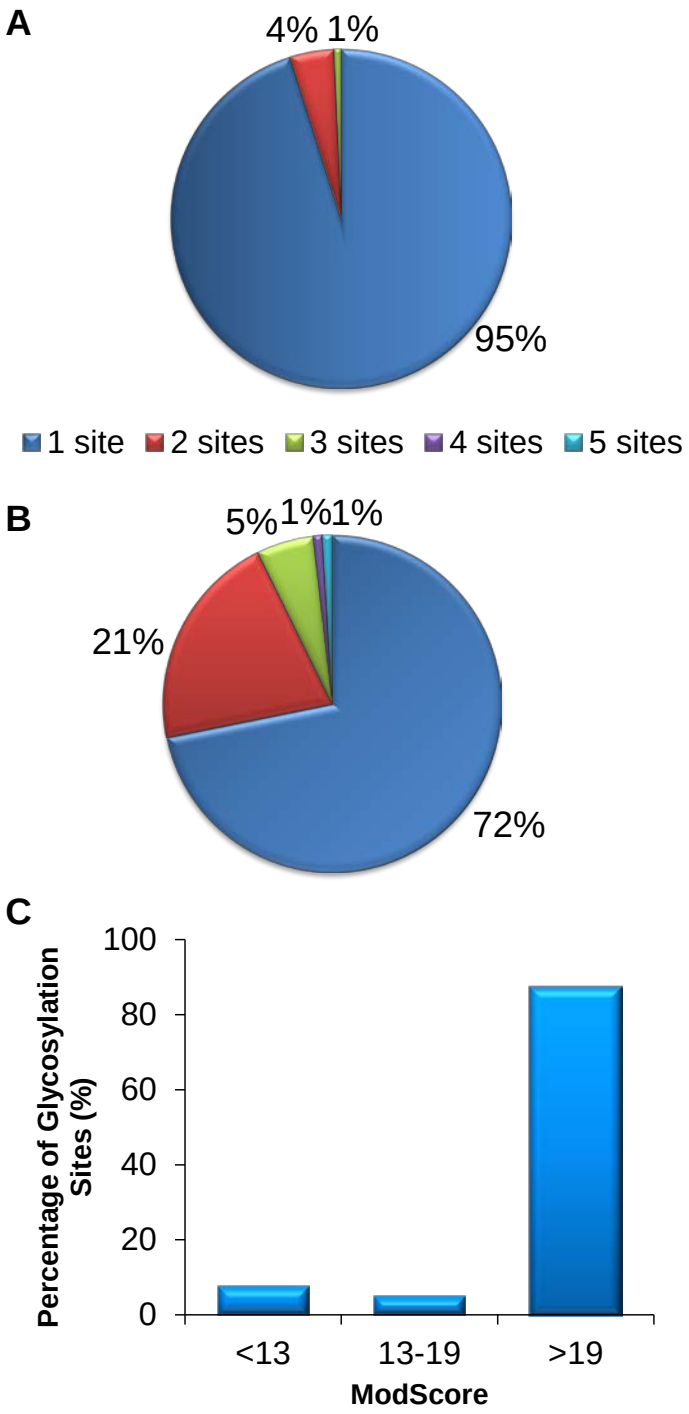

Figure 2.4. Number of N-glycosylation sites identified in (A) peptides and (B) proteins. The ModScore distribution for each site localization is shown in (C).

Overall, 152 sites were identified on 110 cell surface glycoproteins in HEK 293T cells. The major advantage of the current method is that cell surface glycoproteins containing the azido group can be labeled under mild physiological conditions. The limitations of the current method include that the metabolic labeling method is not suitable for clinical tissue samples, but it can be employed for cultured cells and also model animals such as mice and zebrafish.[23, 32] In addition, this method can only be used to identify N-glycopeptides containing GalNAc, but even with this limitation, more than 150 unique glycosylation sites were identified in over one hundred cell surface proteins. The number of glycosylated proteins identified with the current method is comparable to previous results, in which 110 glycoproteins in Jurkat T-lymphocyte cells were identified with the cell surface-capturing method.[9]

The majority of N-glycopeptides identified (95%) contained only one glycosylation site. Figure 2.4A shows the distribution of glycosylation sites in peptides; very few (6 peptides) have two sites, and only one contains three N-glycosylation sites. Similarly, most proteins (72%) are singly glycosylated, and 21% of them contain two N-glycosylation sites (Figure 2.4B). The protein that contains five glycosylation sites was identified as nicastrin (NCSTN), which is an essential subunit of the gamma-secretase complex that is widely expressed in many different tissues.[33]

Based on the same probability principle of Ascore, ModScore evaluates the confidence associated with each site localization.[18, 34-35] A value greater than 19 indicates a higher than 99% confidence level associated with the site localization, and a score greater than 13 corresponds to a higher than 95% confidence level. Among all the

identified sites, 87% have a ModScore larger than 19, and only a very small portion of them (8%) have a ModScore less than 13, as shown in Figure 2.4C.

2.3.3 Cell surface N-glycoprotein clustering

All glycoproteins identified in this work were clustered using the Database for Annotation, Visualization, and Integrated Discovery (DAVID).[36] Membrane proteins were highly enriched with a *P* value of 1.37E-22. Of the 110 total proteins identified, 104 were categorized as membrane proteins. Six proteins were not defined as membrane proteins in the DAVID analysis, which could be due to incorrect annotations of these proteins, contamination and/or experimental errors. Alternatively, they could be non-membrane proteins that are located on the cell surface through an anchor such as glycophosphatidylinositol (GPI) or other modified lipid groups.

All identified glycoproteins were categorized according to biological process using DAVID. The biological processes with the highest enrichment include cell adhesion, cell motion, structure morphogenesis, transport, interspecies interaction, positive immune regulation, cell projection organization, stimulus response and cell recognition (Figure 2.5A), all of which are consistent with the biological processes associated with cell surface proteins. Glycoproteins were also clustered based on molecular function, and the protein functions that were most highly enriched include signal transducer activity, transporter activity, and carbohydrate binding, as shown in Figure 2.5B. Cell surface proteins are known to participate in these biological functions and processes.[37]

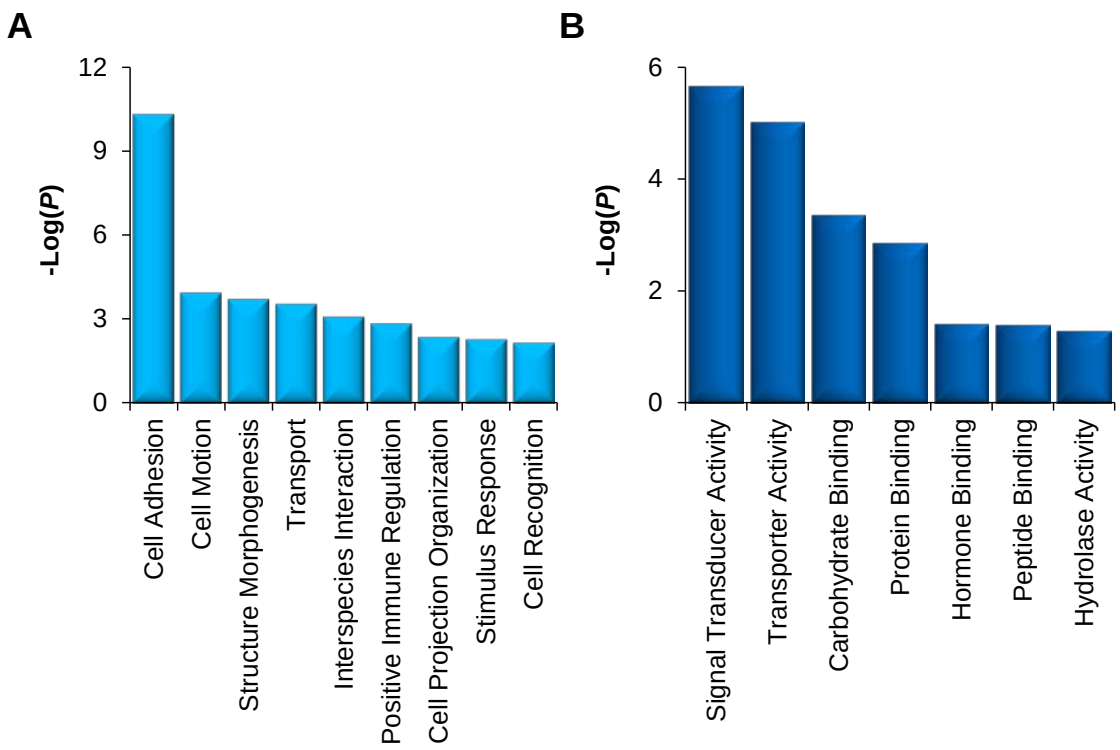

Figure 2.5. Clustering of glycoproteins based on (A) biological process and (B) molecular function.

2.3.4 N-glycosylation sites on cell surface transporters

Membrane transporter proteins compose a large group of cell surface proteins and they regulate the input and output of many molecules and ions. About 10% of human genes are transporter-related, which is corresponds well with the biological significance of transporters and their roles in cell homeostasis.[38] These membrane transporter proteins can serve as potential biomarkers and therapeutic targets. There were 10 transporter proteins identified in this dataset, which are listed in Table 2.1.

Table 2.1. N-glycosylation sites in transporter proteins identified in the current cell surface experiment (# denotes glycosylation site).

Protein	Site	ModScore	Peptide	XCorr	ppm	Annotation
SLC1A4	201 206	1000 1000	VVTQN#SSSGN#VTHEK	3.4	-0.45	Neutral amino acid transporter; transports alanine, serine, cysteine and threonine
SLC1A5	212	1000	SYSTTYEERN#ITGTR	3.2	-0.11	Neutral amino acid transporter; accepts all neutral amino acids
SLC2A1	45	1000	VIEEFYN#QTWVHR	3.6	0.38	Facilitated glucose transporter member
SLC30A6	352	17.0	DDWIRPALLSGPVAANVLN#FSD HHVIPMPLLK	3.0	-0.94	Zinc transporter
SLC39A10	339	106.2	DLNEDDHHHECLN#VTQLLK	3.7	-0.08	Zinc transporter
SLC39A6	266	23.6	NTNENPQECFN#ASK	2.8	0.65	Zinc-influx transporter
	283	24.6	LLTSHGMGIQVPLN#ATEFNYLC PAIINQIDAR	5.0	-0.55	
	67	24.4	YGEN#NSLSVEGFRK	3.8	0.23	
SLC3A2	365	1000	DASSFLAEWQN#ITK	3.1	0.19	Involved in sodium-independent neutral amino acid transport including phenylalanine, tyrosine, leucine, arginine and tryptophan
	381	181.2	LLIAGTN#SSDLQQILSLLESNK	5.3	-0.46	
	424	59.3	SLVTQYLN#ATGNR	3.5	-0.75	
SLC44A2	187	1000	GVLMVGN#ETTYEDGHGSRK	3.5	3.24	Choline transporter-like protein
	417	84.6	TCNPETFPSSN#ESR	2.5	1.76	
SLC4A7	776	13.4	LTSYSCVCTEPPNPSN#ETLAQW K	3.8	3.43	Sodium bicarbonate cotransporter
SLC6A15	187	1000	N#ASHTFVEPECEQSSATTYYW YR	5.0	0.48	Sodium-dependent neutral amino acid transporter, specifically branched-chain amino acids

The protein solute carrier family 3 member 2 (SLC3A2), also designated as cluster of differentiation 98 (CD98), is the 4F2 cell-surface antigen heavy chain and is required for the function of light chain amino-acid transporters.[39] In this experiment, three N-glycopeptides were confidently identified from SLC3A2, *i.e.* DASSFLAEWQN#ITK, LLIAGTN#SSDLQQILSLLESNK and SLVTQYLN#ATGNR with corresponding N-glycosylation sites at N365, N381 and N424.

N-glycosylation may play determinant roles in trafficking these proteins to the cell surface since glycosylation has been proven to regulate the classical secretory pathway.[40] Glycans within proteins can significantly increase the overall hydrophilicity and impact protein interactions with other small molecules or macromolecules. Reversible glycan binding on these cell surface transporters may also participate in the regulation of molecular transport through the cell membrane.

2.3.5 N-glycosylation sites on cluster of differentiation proteins

Molecules presented on the cell surface can differentiate and classify the cell type. Cluster of differentiation (CD) molecules are a group of cell surface molecules that are selected to distinguish cell type. Traditionally these CDs are targets for immunophenotyping and often have a wide variety of functions.[41] For example, some CDs are receptors or ligands and some contribute to cell adhesion. To date, around 450 CDs have been designated to different types of human cells, among which about 380 are surface proteins. In this experiment, 35 N-glycosylation sites were located on 24 CDs in HEK 293T cells and some selected sites are listed in Table 2.2. All sites identified in CDs are listed in a supplemental table online at doi.org/10.1007/s13361-014-1016-7.

Table 2.2. Selected CD proteins identified and their corresponding N-glycosylation sites (# denotes glycosylation site).

Protein	CD	Site	ModScore	Peptide	XCorr	ppm	Annotation
ATP1B3	CD298	124	1000	N#LTVCPDGALFEQK	3.1	-0.12	Na/K-transporting ATPase subunit beta-3
BSG	CD147	160	1000	ILLTCSLN#DSATEVTGHR	2.3	0.17	Basigin; participates in the targeting of monocarboxylate transporters to the plasma membrane
		268	1000	ALMN#GSESR	2.0	-0.11	
CD276	CD276	104	1000	TALFPDLLAQGN#ASLR	4.0	-0.42	May be involved in regulation of T-cell-mediated immune response
		215	1000	VVLGAN#GTYSCLVR	3.4	0.86	
CD46	CD46	83	1000	GYFYIPPLATHTICDRN#HTWLPVSDDACYR	3.1	1.69	Membrane cofactor protein; cofactor for complement factor I
CD97	CD97	371	1000	N#VTMGQSSAR	1.4	0.13	Receptor possibly involved in adhesion and signaling after leukocyte activation
INSR	CD220	445	1000	HN#LTITQGK	2.2	0.06	Insulin receptor; receptor tyrosine kinase that mediates the action of insulin
ITGA1	CD49a	840	1000	FN#VSLTVK	1.7	0.09	Integrin alpha-1; receptor for laminin and collagen
ITGA2	CD49b	105	68.0	LNLQTSTSIPN#VTEMK	3.5	0.99	Integrin alpha-2; receptor for laminin, collagen, collagen C-propeptides, fibronectin and E-cadherin
ITGA5	CD49e	182	1000	EPLSDPVGTCYLSTDN#FTR	4.2	-1.51	Integrin alpha-5; receptor for fibronectin and fibrinogen
MCAM	CD146	56	1000	CGLSQSQGN#LSHVDWFSVHK	3.5	0.19	Cell surface glycoprotein MUC18; participates in cell adhesion

Integrins are cell adhesion receptors that are conserved among species, and play critical roles in developmental and pathological processes. The integrin family is heavily involved in mediating the attachment of cells to the extracellular matrix and also takes part in specialized cell-cell interactions.[42] Integrin alpha-1 (ITGA1), designated as CD49a, acts as a receptor for laminin and collagen and participates in anchorage-dependent, negative regulation of epidermal growth factor-stimulated cell growth. Integrin alpha-2 (ITGA2), CD49b, is a receptor for laminin, collagen, collagen C-propeptides, fibronectin and E-cadherin. It is also accountable for adhesion of platelets and other cells to collagen, the modulation of collagen and collagenase gene expression, and organization of newly synthesized extracellular matrix. Integrin alpha-5 (ITGA5), CD49e, is a receptor for fibronectin and fibrinogen. Glycosylation sites were identified on each of these three cell surface integrins in HEK 293T cells.

CD molecules are crucial for classifying cells, but their glycosylation remains largely unstudied. Considering the importance of these molecules and protein glycosylation on the cell surface, information regarding CD glycosylation can provide insight into the functions of these molecules. Cell surface protein glycosylation analysis can offer valuable information that will lead to the identification of potential drug targets and biomarkers for diseases.

2.4 Conclusions

Glycoproteins on the cell surface are essential for a wide range of cellular events, and their global analysis is exceptionally challenging. A novel method has been developed that integrates metabolic labeling, copper-free click chemistry and mass

spectrometry-based proteomics to comprehensively and site-specifically analyze cell surface N-glycoproteins. Labeling proteins with the azido functional group allows the selective enrichment and separation of cell surface glycoproteins, while the copper-free click chemistry reaction tags cell surface glycoproteins with biotin under physiological conditions. In this experiment, 152 N-glycosylation sites were identified in 110 cell surface proteins in HEK 293T cells by MS. The main functions assigned to identified surface N-glycoproteins, including signal transducer, transporter, binding, and catalytic activity, are consistent with the documented functions of cell surface proteins. The current strategy provides an effective method for large-scale analysis of the cell surface N-glycoproteome, and can be extensively applied for further cell surface studies.

2.5 References

1. Apweiler, R.; Hermjakob, H.; Sharon, N., On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Bba-Gen Subjects* **1999,** *1473* (1), 4-8.

2. Zhao, Y. Y.; Takahashi, M.; Gu, J. G.; Miyoshi, E.; Matsumoto, A.; Kitazume, S.; Taniguchi, N., Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci* **2008,** *99* (7), 1304-1310.

3. Rudd, P. M.; Elliott, T.; Cresswell, P.; Wilson, I. A.; Dwek, R. A., Glycosylation and the immune system. *Science* **2001,** *291* (5512), 2370-2376.

4. Gilgunn, S.; Conroy, P. J.; Saldova, R.; Rudd, P. M.; O'Kennedy, R. J., Aberrant PSA glycosylation-a sweet predictor of prostate cancer. *Nat Rev Urol* **2013,** *10* (2), 99-107.

5. Vigerust, D. J., Protein glycosylation in infectious disease pathobiology and treatment. *Cent Eur J Biol* **2011,** *6* (5), 802-816.

6. Stastna, M.; Van Eyk, J. E., Secreted proteins as a fundamental source for biomarker discovery. *Proteomics* **2012,** *12* (4-5), 722-735.

7. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M., Drug-target network. *Nat Biotechnol* **2007,** *25* (10), 1119-1126.

8. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat Rev Drug Discov* **2002,** *1* (9), 727-730.

9. Wollscheid, B.; Bausch-Fluck, D.; Henderson, C.; O'Brien, R.; Bibel, M.; Schiess, R.; Aebersold, R.; Watts, J. D., Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* **2009,** *27* (4), 378-386.

10. Chang, P. V.; Prescher, J. A.; Hangauer, M. J.; Bertozzi, C. R., Imaging cell surface glycans with bioorthogonal chemical reporters. *J. Am. Chem. Soc.* **2007,** *129* (27), 8400-+.

11. Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute-solution in the presence of detergents and lipids. *Anal. Biochem.* **1984,** *138* (1), 141-143.

12. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

13. Peng, J. M.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J Proteome Res* **2003,** *2* (1), 43-50.

14. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007,** *4* (3), 207-214.

15. Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* **2010,** *143* (7), 1174-1189.

16. Du, X.; Callister, S. J.; Manes, N. P.; Adkins, J. N.; Alexandridis, R. A.; Zeng, X.; Roh, J. H.; Smith, W. E.; Donohue, T. J.; Kaplan, S.; Smith, R. D.; Lipton, M. S., A computational strategy to analyze label-free temporal bottom-up proteomics data. *J Proteome Res* **2008,** *7* (7), 2595-2604.

17. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007,** *4* (11), 923-925.

18. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **2006,** *24* (10), 1285-1292.

19. Asano, N., Naturally occurring iminosugars and related compounds: Structure, distribution, and biological activity. *Curr. Top. Med. Chem.* **2003,** *3* (5), 471-484.

20. Zaro, B. W.; Yang, Y. Y.; Hang, H. C.; Pratt, M. R., Chemical reporters for fluorescent detection and identification of O-GlcNAc-modified proteins reveal glycosylation of the ubiquitin ligase NEDD4-1. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108* (20), 8146-8151.

21. Hsu, T. L.; Hanson, S. R.; Kishikawa, K.; Wang, S. K.; Sawa, M.; Wong, C. H., Alkynyl sugar analogs for the labeling and visualization of glycoconjugates in cells. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (8), 2614-2619.

22. Laughlin, S. T.; Bertozzi, C. R., Metabolic labeling of glycans with azido sugars and subsequent glycan-profiling and visualization via Staudinger ligation. *Nat. Protoc.* **2007,** *2* (11), 2930-2944.

23. Laughlin, S. T.; Baskin, J. M.; Amacher, S. L.; Bertozzi, C. R., *In vivo* imaging of membrane-associated glycans in developing zebrafish. *Science* **2008,** *320* (5876), 664-667.

24. Boyce, M.; Carrico, I. S.; Ganguli, A. S.; Yu, S. H.; Hangauer, M. J.; Hubbard, S. C.; Kohler, J. J.; Bertozzi, C. R., Metabolic cross-talk allows labeling of O-linked beta-N-

acetylglucosamine-modified proteins *via* the N-acetylgalactosamine salvage pathway. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108* (8), 3141-3146.

25. Sletten, E. M.; Nakamura, H.; Jewett, J. C.; Bertozzi, C. R., Difluorobenzocyclooctyne: Synthesis, reactivity, and stabilization by beta-cyclodextrin. *J. Am. Chem. Soc.* **2010,** *132* (33), 11799-11805.

26. Mbua, N. E.; Guo, J.; Wolfert, M. A.; Steet, R.; Boons, G. J., Strain-promoted alkyne-azide cycloadditions (SPAAC) reveal new features of glycoconjugate biosynthesis. *ChemBioChem* **2011,** *12* (12), 1911-1920.

27. Wang, Q.; Chan, T. R.; Hilgraf, R.; Fokin, V. V.; Sharpless, K. B.; Finn, M. G., Bioconjugation by copper(I)-catalyzed azide-alkyne [3+2] cycloaddition. *J. Am. Chem. Soc.* **2003,** *125* (11), 3192-3193.

28. Kaji, H.; Saito, H.; Yamauchi, Y.; Shinkawa, T.; Taoka, M.; Hirabayashi, J.; Kasai, K.; Takahashi, N.; Isobe, T., Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* **2003,** *21* (6), 667-672.

29. Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R., Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* **2003,** *21* (6), 660-666.

30. Agarwal, A. K.; Gude, R. P.; Kalraiya, R. D., Regulation of melanoma metastasis to lungs by cell surface lysosome associated membrane protein-1 (Lamp1) *via* galectin-3. *Biochem. Biophys. Res. Commun.* **2014,** *449* (3), 332-337.

31. Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008,** *7* (1), 29-34.

32. Dube, D. H.; Prescher, J. A.; Quang, C. N.; Bertozzi, C. R., Probing mucin-type O-linked glycosylation in living animals. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (13), 4819-4824.

33. Hu, Y.; Fortini, M. E., Different cofactor activities in gamma-secretase assembly: Evidence for a nicastrin-Aph-1 subcomplex. *J. Cell Biol.* **2003,** *161* (4), 685-690.

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

35. Chen, W. X.; Smeekens, J. M.; Wu, R. H., A universal chemical enrichment method for mapping the yeast N-glycoproteome by MS. *Mol. Cell. Proteomics* **2014,** *13* (6), 1563-1572.

36. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4* (1), 44-57.

37. Lodish, H., *Molecular Cell Biology*. 4th ed.; W. H. Freeman: New York, 2008.

38. Hediger, M. A.; Clemencon, B.; Burrier, R. E.; Bruford, E. A., The ABCs of membrane transporters in health and disease (SLC series): Introduction. *Mol. Asp. Med.* **2013,** *34* (2-3), 95-107.

39. Mastroberardino, L.; Spindler, B.; Pfeiffer, R.; Skelly, P. J.; Loffing, J.; Shoemaker, C. B.; Verrey, F., Amino-acid transport by heterodimers of 4F2hc/CD98 and members of a permease family. *Nature* **1998,** *395* (6699), 288-291.

40. Varki, A., Biological roles of oligosaccharides - All of the theories are correct. *Glycobiology* **1993,** *3* (2), 97-130.

41. Zola, H.; Swart, B.; Banham, A.; Barry, S.; Beare, A.; Bensussan, A.; Boumsell, L.; Buckley, C. D.; Buhring, H. J.; Clark, G.; Engel, P.; Fox, D.; Jin, B. Q.; Macardle, P. J.; Malavasi, F.; Mason, D.; Stockinger, H.; Yang, X. F., CD molecules 2006 - Human cell differentiation molecules. *Journal of Immunological Methods* **2007,** *319* (1-2), 1-5.

42. Barczyk, M.; Carracedo, S.; Gullberg, D., Integrins. *Cell Tissue Res.* **2010,** *339* (1), 269-280.

CHAPTER 3. QUANTIFICATION OF THE CELL SURFACE N-GLYCOPROTEOME THROUGHOUT EPITHELIAL-MESENCHYMAL TRANSITION

3.1 Introduction

Epithelial cells can transform into mesenchymal cells in a process known as the epithelial-mesenchymal transition (EMT). This process is integral for a wide variety of necessary physiological processes, including development and healing, as well as adaptive processes such as tumor progression and metastasis.[1-3] EMT occurs under a variety of conditions, and three types of EMT have been proposed. Type 1 EMTs are involved in embryo formation and organ development, and do not cause systemic invasiveness. Type 2 EMTs are associated with inflammation, such as wound healing and tissue regeneration, and type 3 EMTs are involved in tumor cell metastasis and cancer progression.[4] Despite their differences, cells undergoing EMT appear to have similar processes and characteristics.

Epithelial cells form monolayers that are held together through cell adhesion molecules like E-cadherin, and are attached to a matrix through basement membrane anchors.[5] As cells transition from epithelial to mesenchymal, they lose cell adhesion properties and gain motility and invasive properties.[4] This mesenchymal phenotype allows tumor cells to detach from surrounding cells and enter the circulatory system during metastasis.[6-7]

Since its initial discovery in 1995, EMT has been the focus of many developmental and disease studies, to attain a better understanding of the underlying signaling mechanisms involved in EMT.[8-11] There has been great interest in deciphering the alternate roles EMT plays in embryonic development and wound healing, compared to pathological processes such as fibrosis, and cancer progression.[12-14] As a result, many transcription factors and genes have been found to play a role in EMT, including Snail1, Snail2, Zeb2, and Twist.[5, 12, 15-16] Several hallmarks of this transition have been identified, for example, E-cadherin is down-regulated as cells lose epithelial properties, and N-cadherin is up-regulated as they gain mesenchymal properties. Other proteins up-regulated during EMT are fibronectin, vimentin and laminin.[5] However, surface glycoproteins have yet to be globally analyzed in the context of EMT.

A recent review outlines epigenetic and post-translational modifications that play a role in EMT, including DNA methylation, histone modifications, hydroxylation, phosphorylation, small ubiquitin-like modifier conjugation (SUMOylation) and O-glycosylation.[17] Most of these modifications impact individual proteins involved in EMT, for example, phosphorylation and O-GlcNAcylation of SNAIL both regulate EMT. Additionally, the SUMOylation of FOXM1 represses miR-200b/c in breast cancer cells, and promotes EMT.[18] However, global analysis of these modifications during EMT has yet to be performed.

Another study has investigated how aberrant glycosylation through the hexosamine biosynthesis pathway (HBP) is induced in EMT. The HBP generates UDP-GlcNAc, which is a precursor for O-GlcNAcylation and N-glycosylation. Cell surface glycans were found to have increased α2-6 sialylation, poly-LacNAc, and fucosylation, and O-

GlcNAc was found to be modulated during EMT.[17] Glycoproteins on the cell surface regulate nearly every cellular activity and are known to be highly involved in cell adhesion, motility and invasiveness.[19-20] Based on changes in cell properties during EMT, surface glycoproteins may also play a critical role in EMT and are predicted to be modulated throughout the transition.

In order to investigate cell surface N-glycoprotein changes throughout EMT, we induced EMT in MCF 10A cells with transforming growth factor-β (TGF-β), and paired quantitative multiplexed proteomics techniques with our cell surface glycoprotein analysis method (described in Chapter 2). Our method integrates metabolic labeling, copper-free click chemistry and MS-based proteomics to selectively and site-specifically identify surface N-glycoproteins.[21] Here, through combination with multiplexed proteomics, we systematically quantified cell surface N-glycoproteins throughout EMT. Whole proteome changes were also quantitated with HPLC fractionation and TMT labeling. Systematic investigation of cell surface glycoproteins will provide insight into the molecular mechanisms of EMT.

3.2 Experimental Methods

3.2.1 Cell culture and TGF-β treatment

MCF 10A cells were purchased from ATCC and grown in Mammary Epithelial Cell Growth Medium (MEGM, Lonza) containing bovine pituitary extract (BPE), hydrocortisone, human epidermal growth factor (hEGF), insulin (from the Lonza MEGM BulletKit, CC-3150) and 100 ng/mL cholera toxin.

EMT was induced by treatment with 4 ng/mL TGF-β. All cells were labeled with 100 μM GalNAz for 4 days prior to harvesting. Cells were washed with PBS and media was replaced every two days to ensure consistent concentrations of TGF-β and GalNAz. Different time points (i.e. 0, 4 and 8 days after treatment) were chosen to investigate cell surface glycoproteins throughout EMT. For each time point, when cells were ~90% confluent, they were washed with PBS and the copper-free click reaction was performed with 100 μM dibenzocyclooctyne (DBCO)-sulfo-biotin in cell stripper, for one hour at 37 °C. Cells were harvested by scraping and centrifugation at 300 g for 5 minutes, and washed with 10 mM dithiothreitol (DTT).

3.2.2 Lysis, protein digestion and NeutrAvidin enrichment

Cells were resuspended in digitonin buffer (150 mM NaCl, 25 μg/mL digitonin, 50 mM HEPES, pH=7.9, and protease inhibitor cocktail (1 tablet per 10 mL buffer, Roche)) and incubated 10 minutes at 4 °C with end-over-end rotation. Samples were centrifuged at 2000 g for 10 minutes and supernatants were discarded. Cell pellets were resuspended in buffer (150 mM NaCl, 0.5% sodium deoxycholate (SDC), 20 units/mL benzonase, 50 mM HEPES, pH=7.9, and protease inhibitor cocktail (same concentration as above)) and incubated 45 minutes at 4 °C with end-over-end rotation. Samples were centrifuged at 25,830 g for 10 minutes, and pellets were discarded. Protein concentrations were measured with BCA protein assays.

Protein reduction and alkylation was carried out with 5 mM DTT (30 minutes, 56 °C) and 14 mM iodoacetamide (25 minutes, room temperature in the dark), respectively. Proteins were separated through methanol-chloroform precipitation. Briefly, methanol,

chloroform and water were added to one part sample at a 4:1:3 ratio, and centrifuged at 4696 g for 10 minutes. The top layer was removed, the same volume of methanol was added again, and samples were centrifuged at 4696 g for 10 minutes. Supernatant was discarded and protein pellets were dried.

Proteins were resuspended in buffer containing 50 mM HEPES (pH 7.9), 5% (vol/vol) acetonitrile, and 0.1 M urea, and digested with Lys-C (1:200 enzyme:protein ratio) at 31 °C. The next day, 13 µg trypsin was added and samples were incubated 4 hours at 37 °C. Digestions were quenched with trifluoroacetic acid (TFA) until pH was below 2, and centrifuged to discard precipitate. Peptide samples were desalted on a tC18 Sep-Pak cartridge and split for whole proteome (~30 µg) and surface N-glycoprotein analysis (~6 mg).

3.2.3 Surface N-glycoprotein analysis

Cell surface N-glycopeptides tagged with biotin were separated by NeutrAvidin agarose and desalted as described previously (Section 2.2.3). Each sample (0, 4 or 8 day treated) was split in two for duplicate quantitative analysis with TMTsixplex labeling. The 0 day treated samples were labeled with TMT^6-126 and TMT^6-127, 4 day treated samples with TMT^6-128 and TMT^6-129, and 8 day treated samples with TMT^6-130 and TMT^6-131.[22] After quenching the TMT reaction with 5% hydroxylamine, all six channels were combined, dried, and desalted with a tC18 SepPak cartridge. Samples were dried overnight, dissolved in 50 mM NH_4HCO_3 in heavy-oxygen water, and then subjected to deglycosylation with PNGase F. During the three hour reaction, N-glycans were removed from asparagine residues and simultaneously deaminated to form aspartic acid; since this

reaction is carried out in heavy-oxygen water, ^{18}O is incorporated and serves as a mass tag for glycosylation site identification. After 3 hours, the reaction was quenched with formic acid until pH was below 2. Samples were then purified with stage tips and dried for LC-MS/MS analysis.

3.2.4 Whole proteome analysis

Two samples at each time point (0, 4 or 8 day) were used for duplicate quantitative analysis. The same TMT reagents were used as described above (0 day: TMT^6-126 and TMT^6-127, 4 day: TMT^6-128 and TMT^6-129, 8 day: TMT^6-130 and TMT^6-131). Labeling reactions were quenched with 5% hydroxylamine and 1% of each channel was combined and analyzed with LC-MS/MS to determine the correct volumes of each to mix in order to have a 1:1:1:1:1:1 ratio. Corresponding volumes of each channel were mixed, and the combined sample was dried before desalting with a tC18 Sep-Pak cartridge.

Peptides from the proteome sample were fractionated by high pH reversed-phase HPLC, using a 40 minute gradient of 5-55% 10 mM ammonium formate in 90% ACN, pH 10. Twenty fractions were collected, dried, and purified with stage tips before analysis with LC-MS/MS.

3.2.5 LC-MS/MS analysis

Purified glycopeptide samples were resuspended in 5% ACN and 4% TFA, and analyzed on the same microcapillary HPLC and LTQ Orbitrap Elite as described in Section 2.2.4. Here, a 90 minute gradient of 4-17% ACN (containing 0.125% FA) was

used to separate each glycopeptide sample before detection in the LTQ Orbitrap Elite. Each detection cycle included one full MS scan (resolution: 60,000) in the Orbitrap at the Automatic Gain Control (AGC) target of 1 million, followed by up to 15 MS/MS of the most intense ions in the Orbitrap.

Whole proteome samples were analyzed on the same instrument, with varying 90 minute gradients for the 20 fractions. The gradient for the first five fractions was 2-15% ACN (containing 0.125% FA), the second five was 3-17% ACN (containing 0.125% FA), the third five was 4-20% ACN (containing 0.125% FA), and the final 5 was 8-27% ACN (containing 0.125% FA). The same detection cycles as above were used for data-dependent acquisition in the Orbitrap.

3.2.6 Data analysis

Database searching and filtering for glycopeptides was the same as described in Section 2.2.5, with the addition of the fixed modification on lysine and the N-terminus for TMT labeling (+229.162932). Peptides from whole proteome experiments were searched using the same database, with only the modifications oxidation of methionine (+15.9949) and carbamidomethylation of cysteine (+57.0214). For protein analysis, peptides were filtered to a less than 1% FDR, followed by further filtering to achieve less than 1% FDR at the protein level.

3.2.7 Glycosylation site localization

Glycosylation sites were localized with the ModScore algorithm as described in Section 2.2.6.

3.2.8 Quantification analysis

Glycopeptides and peptides were quantified using the TMT reporter ion intensities in the tandem mass spectra. Isotopic information provided by Thermo was used to calibrate the measured intensities. TMT intensities were averaged for each channel, and then normalized accordingly before subsequent analysis. Median peptide ratios for each unique peptide were calculated for each TMT channel, and median protein ratios were calculated from all unique peptides identified within a protein for each TMT channel. Duplicate channels for each sample (126 and 127, 128 and 129, and 130 and 131) were averaged to obtain abundance ratios at each stage of treatment (0, 4 or 8 days, respectively).

3.3 Results and Discussion

3.3.1 Experimental procedure

TGF-β has been shown to induce EMT in MCF 10A cells, which are non-tumorigenic epithelial mammary cells.[23] Different time points throughout EMT were chosen, i.e. cells were harvested after TGF-β treatment for 0, 4 and 8 days. Untreated MCF 10A cells are rounded, but after treatment with TGF-β, cells lose cell adhesion properties and become more elongated in shape.[5] After treatment for 0, 4 or 8 days, cells were labeled with GalNAz for 48 hours, and subjected to a copper-free click reaction under physiological conditions before harvest. After digestion, peptides were split in two for whole proteome and surface glycopeptide analysis. Peptides from whole proteome samples were fractionated by high pH reversed-phase HPLC, and protein abundances throughout the transition were quantified by TMT labeling and LC-MS/MS.

Glycopeptides were enriched with NuetrAvidin agarose and treated with PNGase F in heavy-oxygen water to generate a common tag for glycopeptide and glycosylation site identification. Prior to TMT labeling, each sample was split in two for duplicate quantitative analysis (Figure 3.1).

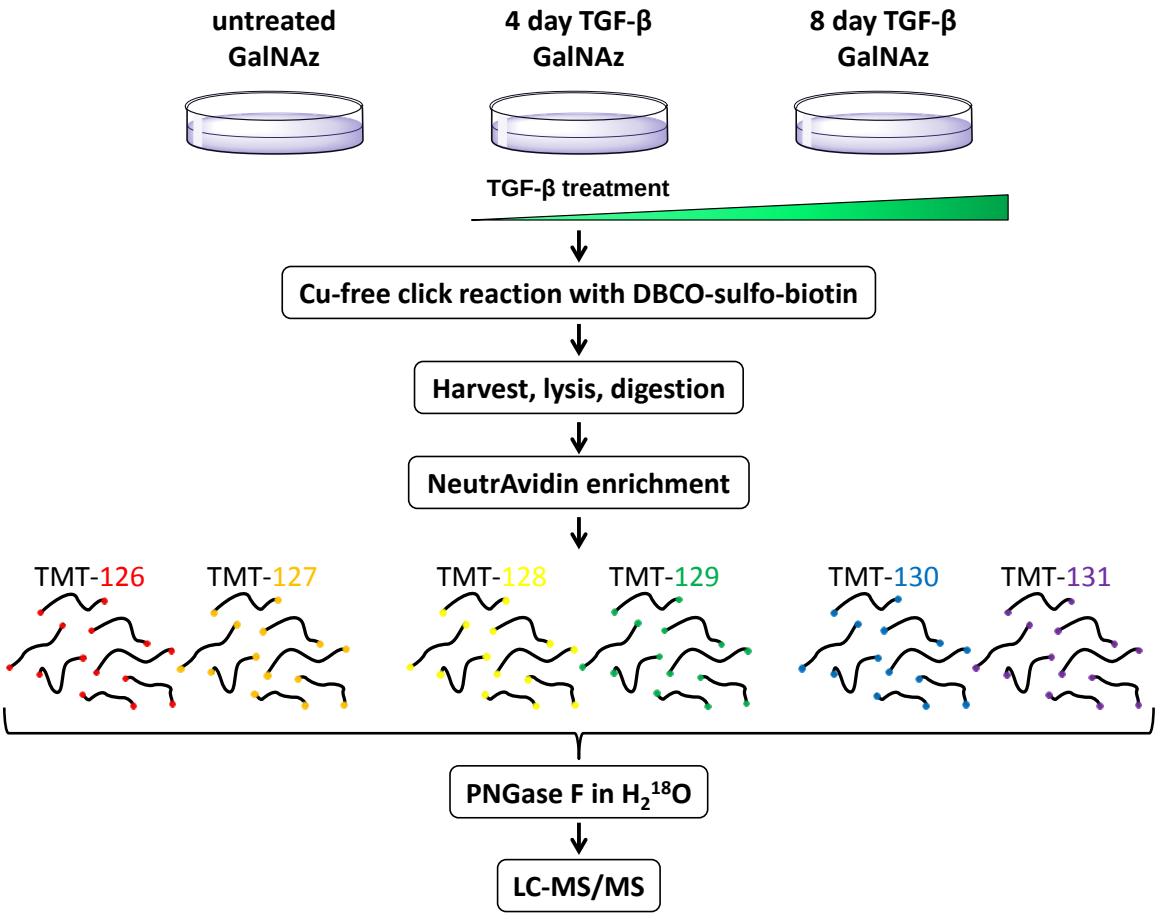

Figure 3.1. Experimental procedure to quantify surface glycoproteins changing throughout EMT.

An example tandem mass spectrum corresponding to the glycopeptide ENQN#HSYSLK (# denotes glycosylation site) is shown in Figure 3.2A. This peptide is from integrin alpha-V (ITGAV), which is a receptor for a variety of proteins, including fibrinogen, laminin and vitronectin. ITGAV is also a subunit of αvβ3 integrin and αvβ6 integrin, which are both up-regulated during EMT and contribute to invasive properties.[24-26] The red box in Figure 3.2A indicates the m/z range where TMT reporter ions are located, which is enlarged in Figure 3.2B. The abundance of this peptide increases after 4 days of treatment, which is consistent with the high expression of ITGAV during EMT, and then decreases slightly at 8 days. It is possible that by day 8, cells have transitioned completely to mesenchymal and therefore the expression of ITGAV is not as high as compared to day 4 during the transition.

3.3.2 Glycopeptide quantification

A total of 438 unique N-glycopeptides corresponding to 235 glycoproteins were quantified here. Duplicate abundances were measured for 0, 4 and 8 day samples, which allowed us to evaluate measurement reproducibility. Figure 3.3 shows the reproducibility between duplicate ratios for glycopeptides quantified after 4 day (Figure 3.3A) and 8 day (Figure 3.3B) treatment. Duplicate intensities from untreated cells were averaged, and intensities from 4 and 8 day cells were each divided by the untreated average to generate the plot. Glycopeptide ratios were very reproducible between duplicates, as shown by the trendline slope of ~0.95 in each plot.

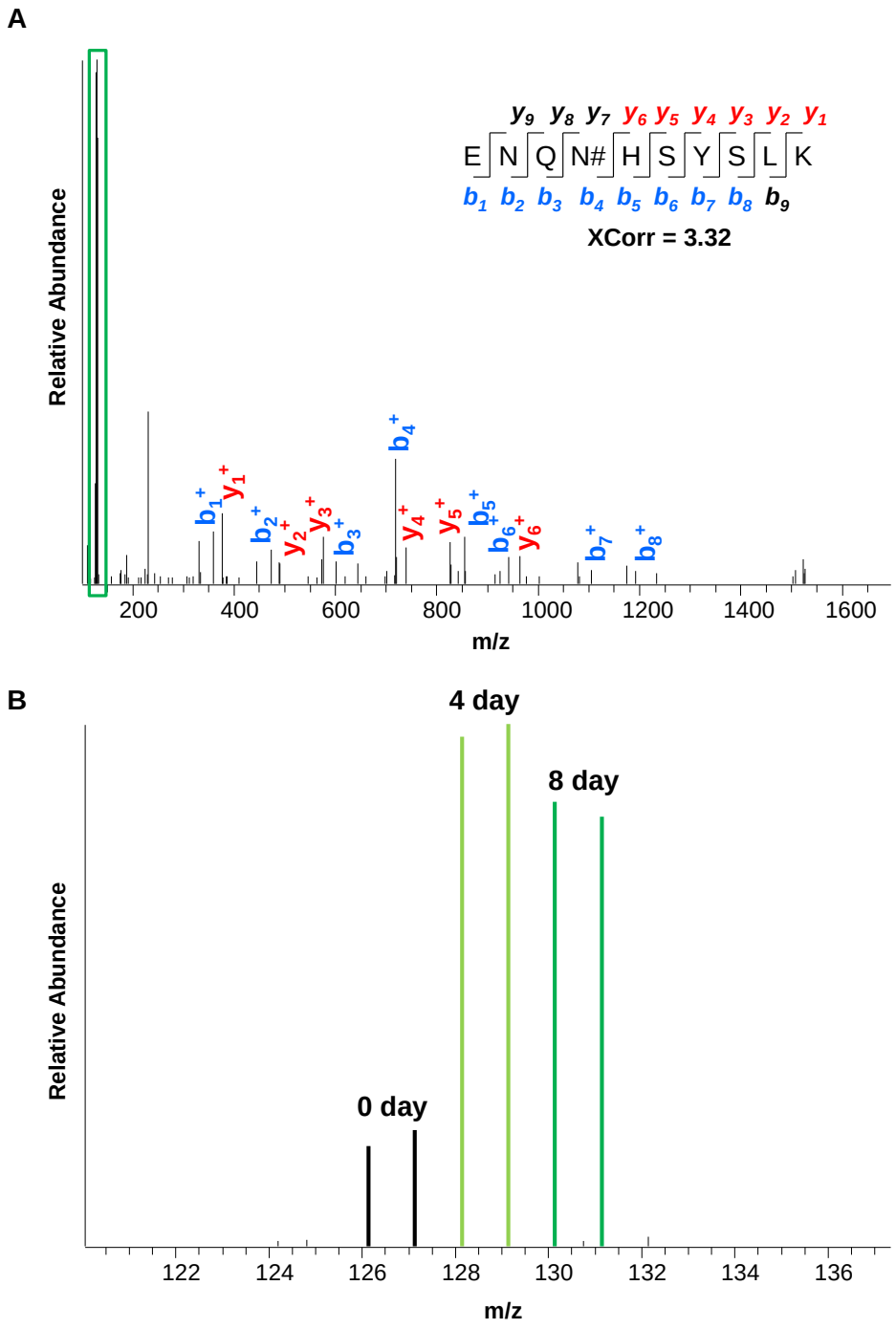

Figure 3.2. Example tandem spectrum for the peptide ENQN#HSYSLK (# denotes glycosylation site) which is from integrin alpha-V, showing (A) peptide fragments and (B) TMT ratios used for quantification. The abundance of this peptide increased after TGF-β treatment for 4 days, and decreased slightly by 8 days of treatment.

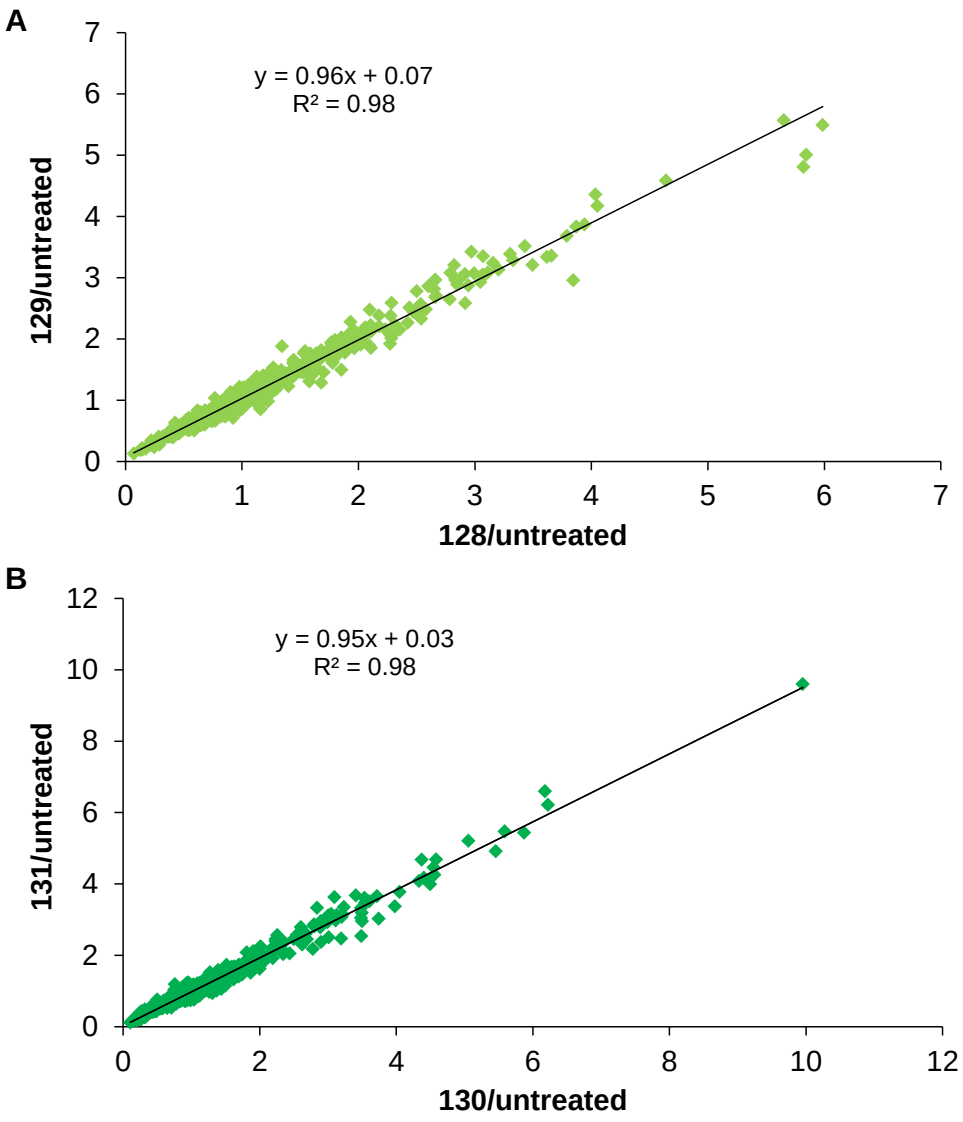

Figure 3.3. Reproducibility of glycopeptide quantification in TMT duplicates for (A) 4 day and (B) 8 day treated versus untreated cells.

The majority of surface N-glycopeptides were neither up- nor down-regulated after 4 or 8 days of treatment with TGF-β, i.e. 351 and 329, respectively, but there were a group of glycopeptides that were up- or down-regulated at each time point. Overall, 39 glycoproteins were down-regulated over the 8 day treatment, and 37 were up-regulated

throughout that time. However, the abundances of 103 glycoproteins increased after 4 days of treatment, which may imply that many glycoproteins are intermediates throughout EMT, and are over-expressed during the transition, and then decrease after the transition is complete. Another potential explanation for the majority of glycoproteins being non-regulated throughout the 8 day treatment is that perhaps the glycans are playing a more dynamic role in the surface properties throughout EMT. Glycopeptides containing only one glycosylation site and a ModScore greater than 13 were normalized to their corresponding protein ratios quantified in the whole proteome experiment, and their adjusted values are shown in Figure 3.4A and 3.4B. After protein normalization, glycopeptide abundance ratios shift towards down-regulation. This shift indicates that glycosylation is regulated as a result of TGF-β treatment, in addition to general protein expression.

Integrins are a group of transmembrane receptor proteins that play critical roles in cell adhesion and extracellular matrix interactions. Four integrins quantified here correlate very well to expression changes that have been previously reported during EMT: ITGA6 was down-regulated, and ITGA5, ITGB1 and ITGB6 were up-regulated (Figure 3.5A, Table 3.1). As mentioned above, αvβ6 integrin has increased expression during EMT, as well as α5β1 integrin, which binds fibronectin, stimulates cell migration and is also up-regulated during EMT.[27-28] ITGA6 was quantified as down-regulated here, and α6β4 integrin has been reported to be down-regulated during EMT.[29] Three integrins quantified here, ITGA10, ITGB5 and ITGB8, have not been reported to be correlated with EMT, and are shown in Figure 3.5B.

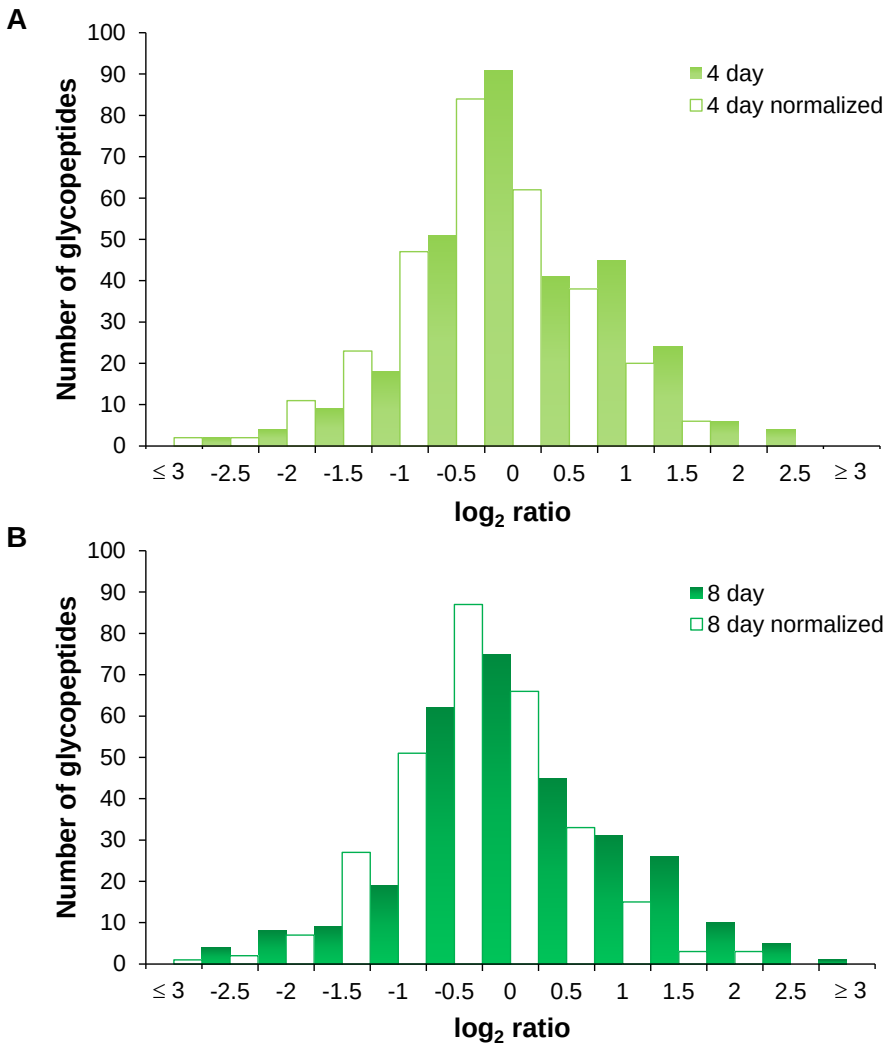

Figure 3.4. Distribution of glycopeptide abundance ratios before and after normalization with protein ratios at (A) 4 and (B) 8 days of TGF-β treatment (bin size is 0.5).

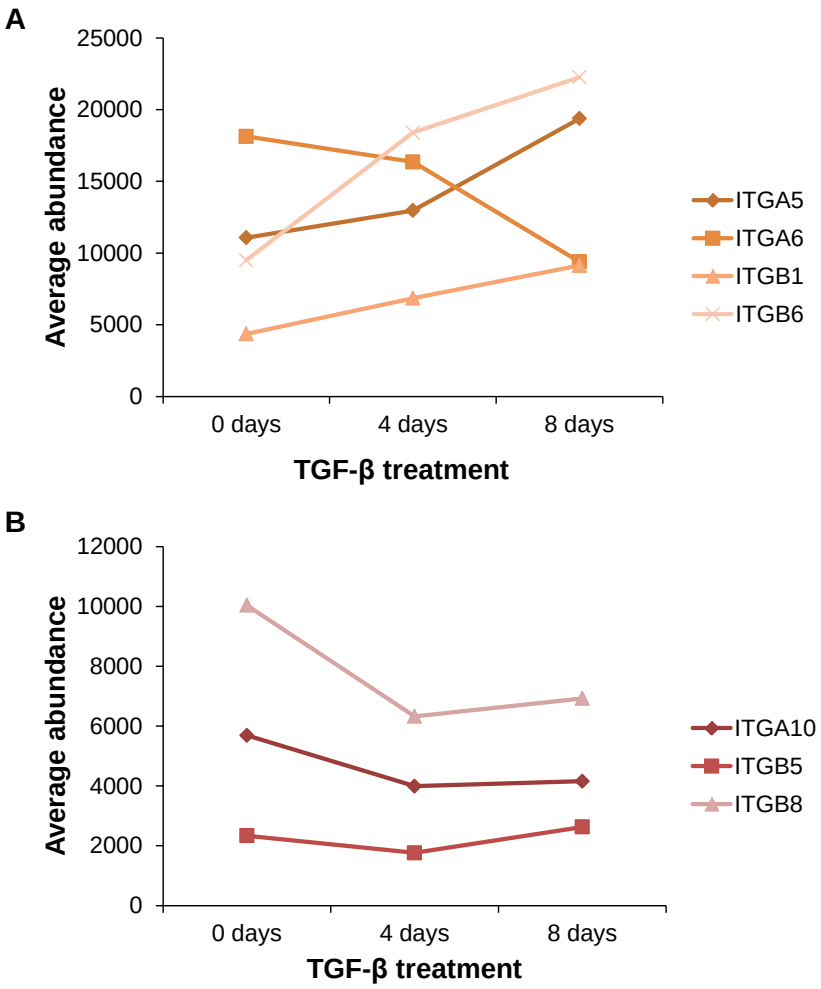

Figure 3.5. Integrins quantified in this work, including those (A) consistent with expression changes previously reported throughout EMT and (B) not previously reported.

Table 3.1. Selected integrins and their corresponding glycosylation site and protein ratios after 4 and 8 days of TGF-β treatment (# denotes glycosylation site).

Gene symbol	Peptide	Site	ModScore	4 day site ratio	4 day protein	4 day normalized	8 day site ratio	8 day protein	8 day normalized	Annotation
ITGA5	TEKEPLSDPVGTCYLSTDN#FTR	182	1000	0.34	1.23	0.28	0.40	1.71	0.23	Integrin alpha-5
	IYLRN#ESEFR	593	1000	0.85		0.69	2.25		1.31	
	NLN#NSQSDVVSFR	773	24.4	1.06		0.86	1.74		1.01	
	VTGLN#CTTNHPINPK	868	31.3	1.49		1.21	2.02		1.18	
ITGA6	AN#HSGAVVLLK	323	1000	1.49	1.62	0.92	0.73	0.97	0.75	Integrin alpha-6
	EINSLN#LTESHNSR	930	73.3	1.19		0.73	0.88		0.91	
	YQTLN#CSVNVNCVNIR	966	97.7	2.73		1.69	0.86		0.89	
ITGB1	NPCTSEQN#CTSPFSYK	212	54.6	1.52	1.47	1.04	1.25	1.42	0.88	Integrin beta-1
	NGVN#GTGENGR	406	73.3	0.55		0.37	0.71		0.50	
	DTCTQECSYFN#ITK	669	1000	4.20		2.86	2.80		1.98	
ITGB6	EVEVN#SSK	463	1000	1.94	3.95	0.49	2.34	4.10	0.57	Integrin beta-6

3.3.3 Whole proteome analysis

Peptides from the whole proteome sample was fractionated into 20 samples with high pH reversed-phase HPLC after digestion, and subsequently analyzed with LC-MS/MS. A total of 4656 proteins were quantified throughout EMT (0, 4 and 8 day treated cells); the distribution of protein abundance ratios after 4 and 8 days of TGF-β treatment are shown in Figure 3.6A. The majority of proteins were not consistently up- or down-regulated over the 8 day treatment; 534 proteins were down-regulated and 699 proteins were up-regulated.

Selected proteins with the largest increase in abundance over 8 days of TGF-β treatment are shown in Figure 3.6B, and those with the largest decrease in abundance over treatment time are in Figure 3.6C. The five most highly up-regulated proteins are implicated in a variety of cellular processes: FTH1, ferritin heavy chain, is important for iron homeostasis, and is also involved in cell proliferation. FTH1 has been shown to be induced during cellular transformation and enhanced by oncogenic signaling.[30] RIN2, Ras and Rab interactor 2, is involved in GTPase activation, and has been previously reported to be a target of the Ras/ERK1/2 pathways which are involved in mesenchymal transformation in pancreatic cancer cells.[31]

The five proteins with the largest decrease in abundance are involved in DNA replication and repair, RNA binding, and translation. HMGN3, high mobility group nucleosome-binding domain-containing protein 3, regulates chromatin structure and related processes such as transcription and DNA repair, and PDS5B, sister chromatid cohesion protein PDS5 homolog B, is involved in DNA replication and repair. RNA-

binding proteins include BRIX1, ribosome biogenesis protein BRX1 homolog, and HNRNPC, heterogeneous nuclear ribonucleoproteins C1/C2. RPL36AL, 60S ribosomal protein L36a-like, is a structural constituent of the ribosome.[32]

3.3.4 Clustering of up- and down-regulated proteins

Up- and down-regulated proteins were each clustered with the Database for Annotation, Visualization and Integrated Discovery (DAVID, version 6.8)[33-34] based on biological process, and the highest enriched processes are shown in Figure 3.7. Processes that were highly enriched among up-regulated proteins include transport and oxidation-reduction process. Interestingly, processes highly enriched among down-regulated proteins include translation initiation and ribonucleoprotein complex biogenesis. Molecular function clustering revealed that down-regulated proteins corresponding to RNA binding are the most highly enriched (P = 7.69E-87), which together implies that translation may be modulated during EMT. There were 215 proteins correlated to RNA binding, a selection of which are shown in Table 3.2. These clustering results are consistent with the functions of the five proteins with the largest decrease in abundance described above, providing further evidence to suggest that translation may be regulated during EMT.

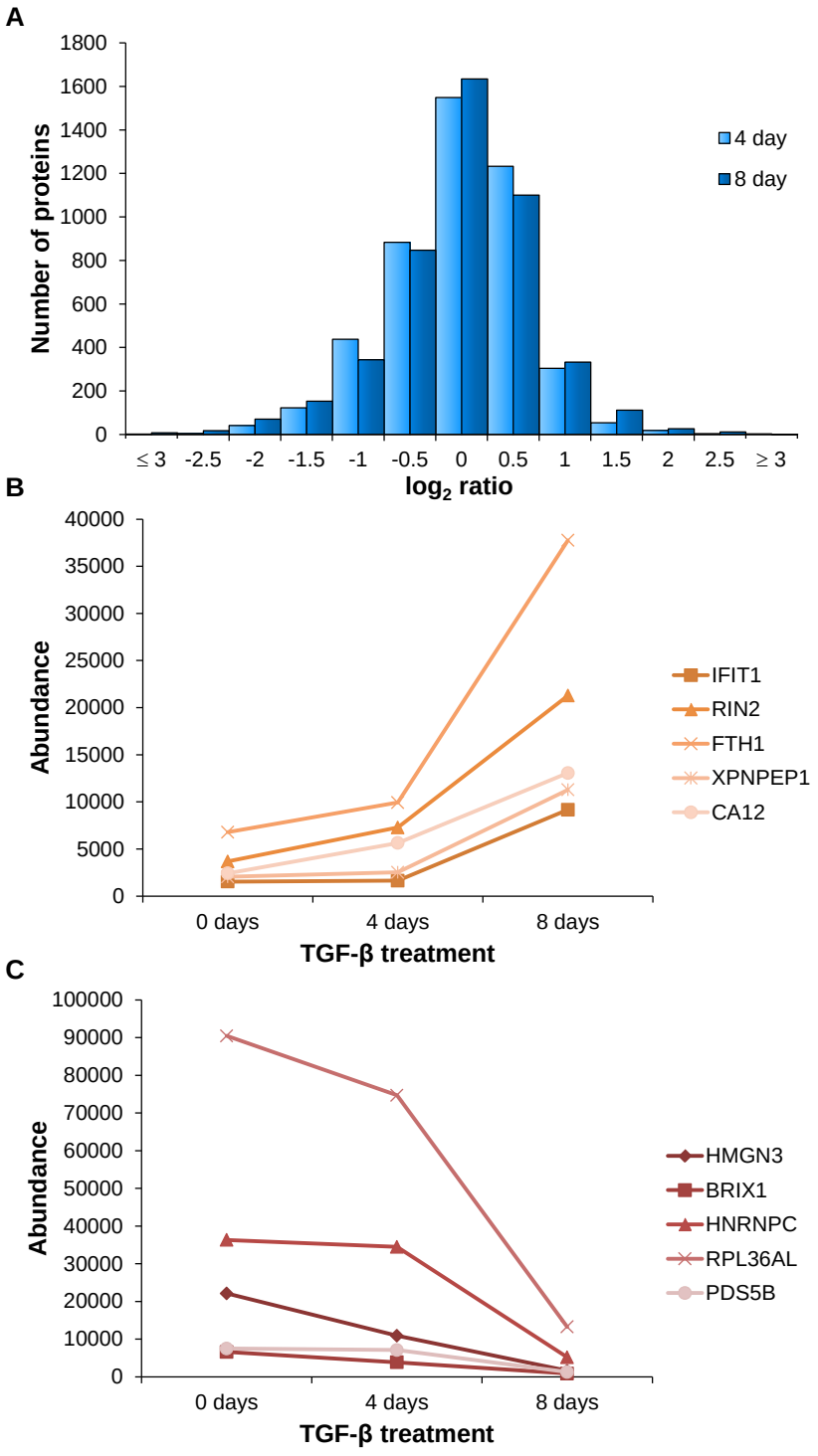

Figure 3.6. Proteome quantification results: (A) protein ratio distribution after 4 and 8 days of TGF-β treatment (bin size is 0.5), (B) examples of proteins with the greatest increase in abundance after 8 days of treatment, and (C) examples of proteins with the greatest decrease in abundance after 8 days of treatment.

Interestingly, several molecular functions were highly enriched in both up- and down-regulated proteins, including cadherin binding and protein binding involved in cell-cell adhesion. The molecular functions listed in DAVID are general and therefore cadherin binding could refer to E-cadherin in down-regulated proteins and N-cadherin in up-regulated proteins. Similarly, cell adhesion related proteins may be highly dynamic throughout EMT, because cells lose their association with epithelial cells and have an increased affinity for mesenchymal cells throughout EMT.[35] Therefore, protein binding involved in cell-cell adhesion can be highly enriched in up- and down-regulated proteins.

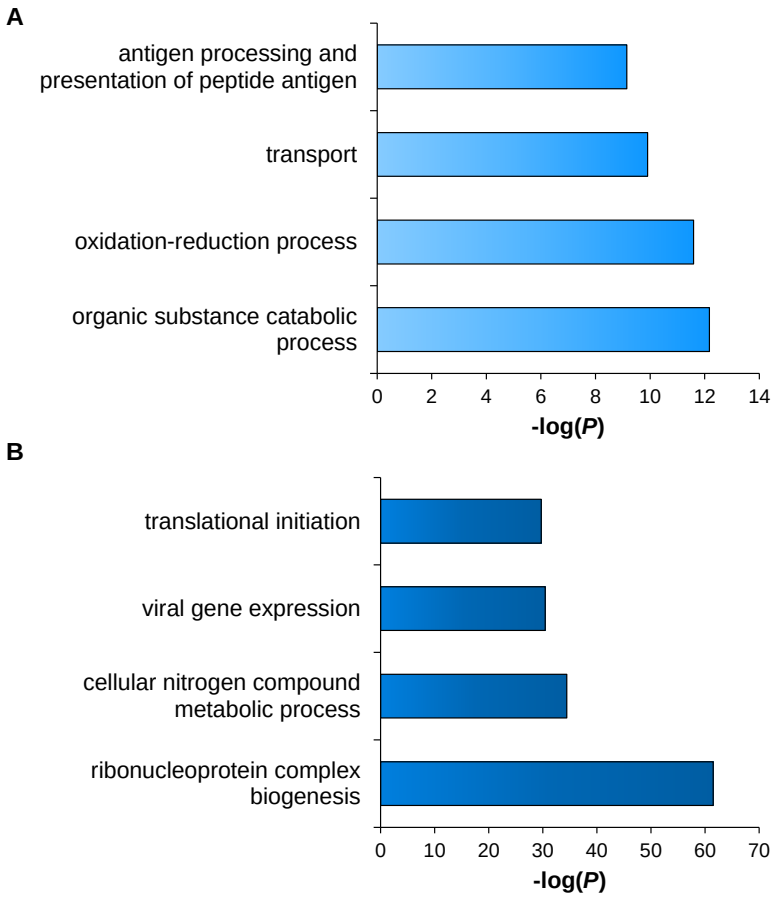

Figure 3.7. Clustering of (A) up- and (B) down- regulated proteins according to biological process.

Table 3.2. Selected down-regulated proteins corresponding to RNA binding (*P* = 7.69E-87).

UniProt Accession	Gene symbol	Number of unique peptides	4 day abundance ratio	8 day abundance ratio	Annotation
Q99700	ATXN2	2	0.51	0.28	Ataxin-2
Q9H0S4	DDX47	4	0.54	0.24	Probable ATP-dependent RNA helicase DDX47
Q9H1E3	NUCKS1	3	0.56	0.22	Nuclear ubiquitous casein and cyclin-dependent kinase substrate 1
Q8TDN6	BRIX1	3	0.58	0.13	Ribosome biogenesis protein BRX1 homolog
Q99848	EBNA1BP2	6	0.58	0.20	Probable rRNA-processing protein EBP2
P63173	RPL38	6	0.58	0.35	60S ribosomal protein L38
P62979	RPS27A	6	0.61	0.18	Ubiquitin-40S ribosomal protein S27a
P62280	RPS11	13	0.62	0.28	40S ribosomal protein S11
Q15024	EXOSC7	2	0.64	0.47	Exosome complex component RRP42
Q8NC51	SERBP1	11	0.65	0.37	Plasminogen activator inhibitor 1 RNA-binding protein
Q9NPE3	NOP10	2	0.66	0.22	H/ACA ribonucleoprotein complex subunit 3

3.4 Conclusions

EMT is an integral process in cells, contributing to development and healing, and can be adapted and exploited for migration and tumor metastasis. Cell surface glycoproteins play critical roles in cell-cell adhesion and extracellular interactions, and are implicated in EMT. Here we have quantitatively and site-specifically investigated surface N-glycoprotein changes throughout EMT by employing metabolic labeling, copper-free click chemistry, and multiplexed proteomics techniques. Overall, 39 glycoproteins were down-regulated and 37 were up-regulated in MCF 10A cells after 8 day treatment with TGF-β. Glycoprotein analysis before and after normalization with corresponding parent protein ratios indicate that glycosylation is indeed modulated throughout EMT, in addition to protein expression. Whole proteome analysis revealed down-regulated proteins to be highly enriched with RNA and translation processes, indicating that translation is modified during EMT. These results provide insight into site-specific surface glycoprotein changes and whole proteome expression modulation as a result of EMT.

3.6 References

1. Shook, D.; Keller, R., Mechanisms, mechanics and function of epithelial-mesenchymal transitions in early development. *Mech Develop* **2003,** *120* (11), 1351-1383.

2. Thiery, J. P.; Acloque, H.; Huang, R. Y. J.; Nieto, M. A., Epithelial-Mesenchymal Transitions in Development and Disease. *Cell* **2009,** *139* (5), 871-890.

3. Xu, J.; Lamouille, S.; Derynck, R., TGF-beta-induced epithelial to mesenchymal transition. *Cell Res* **2009,** *19* (2), 156-172.

4. Kalluri, R.; Weinberg, R. A., The basics of epithelial-mesenchymal transition. *J Clin Invest* **2009,** *119* (6), 1420-1428.

5. Lamouille, S.; Xu, J.; Derynck, R., Molecular mechanisms of epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **2014,** *15* (3), 178-196.

6. Nawshad, A.; LaGamba, D.; Polad, A.; Hay, E. D., Transforming growth factor-beta signaling during epithelial-mesenchymal transformation: Implications for embryogenesis and tumor metastasis. *Cells Tissues Organs* **2005,** *179* (1-2), 11-23.

7. Zavadil, J.; Bottinger, E. P., TGF-beta and epithelial-to-mesenchymal transitions. *Oncogene* **2005,** *24* (37), 5764-5774.

8. Hay, E. D., An overview of epithelio-mesenchymal transformation. *Acta Anat* **1995,** *154* (1), 8-20.

9. Gonzalez, D. M.; Medici, D., Signaling mechanisms of the epithelial-mesenchymal transition. *Sci Signal* **2014,** *7* (344).

10. Gregory, P. A.; Bert, A. G.; Paterson, E. L.; Barry, S. C.; Tsykin, A.; Farshid, G.; Vadas, M. A.; Khew-Goodall, Y.; Goodall, G. J., The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* **2008,** *10* (5), 593-601.

11. Nieto, M. A., The ins and outs of the epithelial to mesenchymal transition in health and disease. *Annu Rev Cell Dev Bi* **2011,** *27*, 347-376.

12. Lim, J.; Thiery, J. P., Epithelial-mesenchymal transitions: insights from development. *Development* **2012,** *139* (19), 3471-3486.

13. O'Connor, J. W.; Gomez, E. W., Biomechanics of TGFβ-induced epithelial-mesenchymal transition: implications for fibrosis and cancer. *Clinical and Translational Medicine* **2014,** *3* (1), 23.

14. Thiery, J. P., Epithelial-mesenchymal transitions in development and pathologies. *Curr Opin Cell Biol* **2003,** *15* (6), 740-746.

15. Hanahan, D.; Weinberg, R. A., Hallmarks of Cancer: The Next Generation. *Cell* **2011,** *144* (5), 646-674.

16. Kurcon, T.; Liu, Z. Y.; Paradkar, A. V.; Vaiana, C. A.; Koppolu, S.; Agrawal, P.; Mahal, L. K., miRNA proxy approach reveals hidden functions of glycosylation. *P Natl Acad Sci USA* **2015,** *112* (23), 7327-7332.

17. Serrano-Gomez, S. J.; Maziveyi, M.; Alahari, S. K., Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Mol Cancer* **2016,** *15*.

18. Wang, C. M.; Liu, R. H.; Wang, L. Z.; Nascimento, L.; Brennan, V. C.; Yang, W. H., SUMOylation of FOXM1B alters its transcriptional activity on regulation of MiR-200 family and JNK1 in MCF7 human breast cancer cells. *Int J Mol Sci* **2014,** *15* (6), 10233-10251.

19. Zhao, Y. Y.; Takahashi, M.; Gu, J. G.; Miyoshi, E.; Matsumoto, A.; Kitazume, S.; Taniguchi, N., Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci* **2008,** *99* (7), 1304-1310.

20. Woods, R. J.; Edge, C. J.; Dwek, R. A., Protein Surface Oligosaccharides and Protein Function. *Nat Struct Biol* **1994,** *1* (8), 499-501.

21. Smeekens, J. M.; Chen, W. X.; Wu, R. H., Mass spectrometric analysis of the cell surface N-glycoproteome by combining metabolic labeling and click chemistry. *Journal of the American Society for Mass Spectrometry* **2015,** *26* (4), 604-614.

22. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C., Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003,** *75* (8), 1895-1904.

23. Zhang, J. Y.; Tian, X. J.; Zhang, H.; Teng, Y.; Li, R. Y.; Bai, F.; Elankumaran, S.; Xing, J. H., TGF-beta-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal* **2014,** *7* (345).

24. Yilmaz, M.; Christofori, G., EMT, the cytoskeleton, and cancer cell invasion. *Cancer Metast Rev* **2009,** *28* (1-2), 15-33.

25. McNiven, M. A., Breaking away: matrix remodeling from the leading edge. *Trends Cell Biol* **2013,** *23* (1), 16-21.

26. Shah, P. P.; Fong, M. Y.; Kakar, S. S., PTTG induces EMT through integrin alpha(V)beta(3)-focal adhesion kinase signaling in lung cancer cells. *Oncogene* **2012,** *31* (26), 3124-3135.

27. Maschler, S.; Wirl, G.; Spring, H.; Bredow, D. V.; Sordat, I.; Beug, H.; Reichmann, E., Tumor cell invasiveness correlates with changes in integrin expression and localization. *Oncogene* **2005,** *24* (12), 2032-2041.

28. Mise, N.; Savai, R.; Yu, H. Y.; Schwarz, J.; Kaminski, N.; Eickelberg, O., Zyxin is a transforming growth factor-beta (TGF-beta)/Smad3 target gene that regulates lung cancer cell motility via integrin alpha 5 beta 1. *J Biol Chem* **2012,** *287* (37), 31393-31405.

29. Yang, X. F.; Pursell, B.; Lu, S. L.; Chang, T. K.; Mercurio, A. M., Regulation of beta 4-integrin expression by epigenetic modifications in the mammary gland and during the epithelial-to-mesenchymal transition. *J Cell Sci* **2009,** *122* (14), 2473-2480.

30. Truitt, M. L.; Conn, C. S.; Shi, Z.; Pang, X. M.; Tokuyasu, T.; Coady, A. M.; Seo, Y.; Barna, M.; Ruggero, D., Differential requirements for eIF4E dose in normal development and cancer. *Cell* **2015,** *162* (1), 59-71.

31. Botta, G. P.; Reginato, M. J.; Reichert, M.; Rustgi, A. K.; Lelkes, P. I., Constitutive K-Ras(G12D) activation of ERK2 specifically regulates 3D invasion of human pancreatic cancer cells via MMP-1. *Mol Cancer Res* **2012,** *10* (2), 183-196.

32. Consortium, T. U., Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **2014,** *42* (D1), D191-D198.

33. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4* (1), 44-57.

34. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009,** *37* (1), 1-13.

35. Theveneau, E.; Mayor, R., Cadherins in collective cell migration of mesenchymal cells. *Curr Opin Cell Biol* **2012,** *24* (5), 677-684.

CHAPTER 4. GLOBAL ANALYSIS OF SECRETED PROTEINS AND GLYCOPROTEINS IN *SACCHAROMYCES CEREVISIAE*

Adapted with permission from American Chemical Society

Smeekens, J.M., Xiao, H., Wu, R. Global analysis of secreted proteins and glycoproteins in *Saccharomyces cerevisiae*. *Journal of Proteome Research*. 2017, 16 (2), 1039-1049.

4.1 Introduction

Protein secretion plays extremely important roles in nearly every extracellular activity, including cell-cell communication, cell-matrix interactions, and cellular immune response.[1] Secreted proteins are often reflective of the developmental and disease status of the cell, including cell differentiation and cancer cell metastasis. Therefore, secreted proteins can provide valuable disease information, and secreted proteins in bodily fluids are highly promising, non-invasive biomarkers for disease detection and surveillance.[2,3]

There are two types of protein secretion pathways in cells: classical and non-classical secretion. Non-classical secretion is very elusive, but several ways have been reported for the secretion of proteins lacking signal peptides.[4-6] In classical secretion, proteins containing an N-terminal signal peptide are synthesized in the ribosome, and translocated to the endoplasmic reticulum (ER), where proteins are glycosylated and folded. They are subsequently transported to the Golgi for further modification, and eventually are sent to the extracellular space by secretory vesicles.[7] Protein N-glycosylation plays extraordinary roles in protein folding and trafficking, frequently determines protein stability, and regulates protein interactions with other proteins or

ligands.[8-12] For instance, the secretion of a group of proteins may be regulated by their glycosylation while glycosylation on other proteins may play different roles, such as regulating their interactions with other molecules or protecting proteins from degradation. However, systematic and quantitative investigation of protein secretion and the correlation with glycosylation remains to be explored. In this work, yeast was used as a model eukaryotic system to systematically study protein secretion, and quantify secreted protein and glycoprotein abundance changes as a result of N-glycosylation inhibition *via* tunicamycin.

Although computational approaches can provide some information regarding protein secretion, specifically by identifying signal peptides, there is still uncertainty regarding the sequences of signal peptides.[13-14] In addition, computational methods have been developed based on experimental results, and often computational results need to be further verified through experimental procedures. Many non-classically secreted proteins are also more difficult to be predicted through computation. Therefore, there is an urgent need to study protein secretion experimentally. It is challenging to systematically investigate secreted proteins and glycoproteins in mammalian cell lines.[15-16] First, it is difficult to distinguish secreted glycoproteins from those in the original culture medium containing fetal bovine serum (FBS). Because of protein conservation, many tryptic peptides from proteins secreted from mammalian cells and those in FBS share the same sequences. Second, highly abundant proteins from FBS (typically 10% in media) prevent the global identification of low-abundance secreted proteins and glycoproteins by current methods. Depleting the media of serum is commonly used to study secreted proteins in

mammalian cells,[17-18] although this alteration may act as a stimulus during cell growth.[19-21]

Mass spectrometry (MS)-based proteomics has made great strides over the past decade, which provides the possibility to globally analyze proteins and their modifications.[22-33] However, due to the complexity of biological samples and wide range of protein abundances, it is very challenging to reach proteins with low abundance.[17, 34-35] Many secreted proteins and glycoproteins fall into this category and their comprehensive identification requires several challenges to be overcome. For example, they are present at much lower abundances compared to the majority of intracellular proteins. Correspondingly, the death of any individual cells during cell growth or subsequent experimental steps results in the release of intracellular proteins into the media where secreted proteins are located. Due to the vast differences in abundances, these intracellular proteins will mask the identification of low-abundance secreted proteins during mass spectrometry analysis. For comprehensive analysis, secreted proteins ideally would be completely isolated from cells, therefore preventing intracellular protein contamination. However, highly effective separation of secreted proteins is not trivial when working with living cells. Centrifugation is the most widely used method for cell harvest, but even common centrifugation speeds may cause cells to break open, therefore releasing intracellular proteins into the media before separation.

Additionally, comprehensive analysis of protein modifications is a daunting task due to their sub-stoichiometry, occurrence on low-abundance proteins and dynamic nature.[36-42] Effective enrichment prior to MS analysis is typically required for complex biological samples, including whole cell lysates and serum samples.[23, 43-47] It is even more

challenging to globally analyze protein glycosylation compared to other types of modifications because glycans are incredibly diverse.[48-57] The high heterogeneity of glycans is incompatible with the typical large-scale identification of protein modifications by MS, which requires a common tag. Furthermore, the variety of glycans also makes the enrichment of glycopeptides much more difficult. Considering that many secreted proteins are glycoproteins that contain relevant disease information, there is a long-standing interest to analyze glycoproteins in the secretome.[16, 58-59] However, comprehensive analysis of protein N-glycosylation in the yeast secretome has yet to be reported.

By combining stable isotope labeling of amino acids in cell culture (SILAC)[60] with MS-based proteomics, we systematically investigated secreted proteins and glycoproteins in yeast. After inhibiting protein N-glycosylation for several hours, we found many secreted proteins were down-regulated, while the abundances of some secreted proteins were relatively unaffected. We also comprehensively and site-specifically studied glycoproteins in the secretome. For the first time, we systematically investigated secreted proteins and glycoproteins with and without the inhibition of protein N-glycosylation in *Saccharomyces cerevisiae*, which can provide a better understanding of protein secretion and glycosylation.

4.2 Experimental Methods

4.2.1 Yeast cell growth, collection of secreted proteins and mixing

Biological triplicate experiments were performed, and for each triplicate, yeast was inoculated overnight in Lys-dropout media (2.0 g/L SCM-Lys, 6.7 g/L Yeast

Nitrogen Base without Amino Acids, 20 g/L dextrose) supplemented with Lys^0. The following day, the overnight culture was diluted into two flasks with 200 mL of Lys-dropout media and 76 mg/L Lys^8 (Heavy) or Lys^0 (Light) until the optical density (OD) at 600 nm was ~0.1. Cells were incubated at 30 °C with shaking at 250 rpm. When the OD was ~0.55, either 2 µg/mL tunicamycin or the same volume (40 µL) of dimethyl sulfoxide (DMSO) was added to heavy or light cells, respectively. After cells were incubated for half an hour, the media was changed by centrifuging cells at 1000 g for 5 min, discarding supernatants and resuspending the cell pellets in new media (Lys-dropout media, plus Lys^8 and tunicamycin, or Lys^0 and DMSO). Cells were incubated for two hours and finally harvested by centrifugation as gently as possible to minimize extra pressure on the cells that would cause cell death and corresponding lysis; acceleration and deceleration rates were decreased to accommodate these concerns. First, cell suspensions were centrifuged at 500 g for 10 minutes. Supernatants were carefully transferred to new tubes and centrifuged again at 1000 g for 10 minutes. Then supernatants were transferred to new tubes and centrifuged again at 4696 g for 10 minutes. Supernatants were transferred again into new tubes, and heavy and light media was mixed based on the OD at 600 nm at the time of tunicamycin or DMSO treatment. The OD is directly correlated to the number of cells per unit volume. The measured ODs were converted into the number of cells in corresponding volumes of cultured media, and then media were mixed based on the same number of cells from each sample. Mixed media was subsequently concentrated on 10 kDa filters until the volume was ~1.0 mL.

4.2.2 Protein reduction, alkylation, precipitation and digestion

Disulfide bonds within proteins were reduced with 5 mM DTT for 25 minutes at 56 °C, and subsequently alkylated with 14 mM iodoacetamide for 30 minutes at room temperature in the dark. Proteins were precipitated with the methanol chloroform precipitation method and the resulting protein pellet was dried in a speed-vacuum concentrator (Labconco).[61] Proteins were resuspended in a buffer with 50 mM HEPES (pH=7.9), 0.1 M urea and 5% ACN, and digested with 70 µg Lys-C overnight at 31 °C with shaking. Digestions were quenched by acidification with trifluoroacetic acid until the pH was ~2. Digests were purified with a 50 mg SepPak tC18 cartridge and separated into two: 10% for protein analysis and 90% for glycoprotein analysis.

4.2.3 Protein analysis

Dried peptides were further purified with the stage-tip method and fractionated into three samples during elution with 20%, 50% or 80% ACN containing 0.1% FA, respectively. Samples were dried and analyzed with an on-line LC-MS/MS system, and different fractions were separated with different LC gradients prior to MS analysis (described below).

4.2.4 Protein N-glycosylation analysis

Dried peptides were enriched with boronic acid conjugated magnetic beads as reported previously.[62] Reversible covalent interactions between glycans and boronic acid were formed under basic conditions. After removing non-glycosylated peptides, glycopeptides were eluted from the beads under acidic conditions. Enriched glycopeptides were dried, and then desalted with a SepPak tC18 cartridge. Purified glycopeptides were further dried overnight in the speed-vacuum concentrator.

Completely dried glycopeptides were incubated with PNGase F in heavy-oxygen water ($H_2{}^{18}O$) for 3 hours at 37 °C, as described previously (Section 2.2.3). Similarly, samples were purified with the stage-tip method, and separated into three fractions with 20%, 50% or 80% ACN containing 0.1% FA. Finally, samples were dried, and then dissolved in a solvent containing 5% ACN and 4% FA for LC-MS/MS analysis.

4.2.5 LC-MS/MS analysis

The same UltiMate HPLC system outlined in previous chapters was used here to fractionate peptides prior to MS analysis. For secreted protein samples, the first fraction was separated using a gradient of 4-13% ACN (with 0.125% FA), the second one using the gradient of 6-30% ACN (with 0.125% FA) and the third was 9-37% ACN (with 0.125% FA). For glycosylation samples, the first fraction was separated using a gradient of 3-22% ACN, the second was 6-30% and the third was 12-45% ACN (with 0.125% FA). Samples were detected in a LTQ Orbitrap Elite, as described previously (Section 2.2.4).

4.2.6 Data analysis

Raw files recorded by MS were converted into mzXML format, and searched using the SEQUEST algorithm (version 28)[63] against a database encompassing sequences of all proteins (6607 protein entries) in the yeast ORFs database downloaded from SGD (http://www.yeastgenome.org/). FDRs of each peptide was estimated, as described previously. The following parameters were used for the database search: 20 ppm precursor mass tolerance; 1.0 Da product ion mass tolerance; Lys-C digestion; up to three missed cleavages; variable modifications: oxidation of methionine (+15.9949) and ^{18}O

tag on Asn (+2.9883) for glycosylation samples; fixed modifications: carbamidomethylation of cysteine (+57.0214). LDA was used to distinguish correct and incorrect peptide identifications, as described previously. For protein glycosylation analysis, the data sets were restricted to only glycopeptides when determining FDRs, which can minimize the false positive identifications of glycopeptides.

4.2.7 Glycosylation site localization

ModScores were calculated to assign glycosylation site localizations and measure the assignment confidence, as described previously. We considered sites with a ModScore >19 ($P < 0.01$) to be confidently localized.

4.2.8 Peptide quantification

Proteins identified based on only one total peptide hit were removed from the datasets to increase the identification confidence. Quantified peptides were filtered based on S/N ratios for both heavy and light peptides. If a peptide (heavy or light) had an S/N ratio less than three, the corresponding peptide (light or heavy, respectively) was required to have an S/N ratio above five, otherwise the peptides were deleted. Peptides identified more than once were quantified by calculating the median value of all peptide area ratios to obtain a final peptide abundance change. Similarly, to quantify protein abundance changes, the median was calculated for all unique peptides for each protein to determine the final protein abundance change.

For glycopeptide identification, after filtering, we manually checked some glycopeptides with relatively low XCorr values, especially for long peptides, and

removed them if their identifications were suspicious. Glycopeptides were quantified in the same way as described above.

4.2.9 Bioinformatic analysis

UniProt designations of extracellular, secreted, and membrane were used to determine which proteins in this dataset were located in these cellular components.[64] DAVID (Database for Database for Annotation, Visualization and Integrated Discovery) was also employed to confirm subcellular locations and to cluster proteins and glycoproteins.[65] SignalP 4.1 was used to calculate which proteins contained a signal peptide sequence.[13]

4.3 Results and Discussion

4.3.1 Experimental procedure

In this work, biological triplicate experiments were performed, which allows the assessment of reproducibility and increases the confidence of the identification and quantification of secreted proteins. In order to systematically investigate protein secretion, tunicamycin was chosen to treat cells because it has been widely used as a potent protein N-glycosylation inhibitor. In this SILAC experiment, heavy cells were treated with tunicamycin and mixed with light cells (treated with DMSO) at a 1:1 ratio. Secretomes were isolated and proteins were reduced, alkylated and precipitated before digestion with Lys-C overnight. Digested peptides were split into two for protein (10%) and glycosylation (90%) analysis, respectively. Glycosylation analysis included enrichment with boronic acid-conjugated magnetic beads, deglycosylation with PNGase

F in heavy-oxygen water, and purification with the stage-tip method before analysis with LC-MS/MS. The experimental procedure for each triplicate is shown in Figure 4.1.

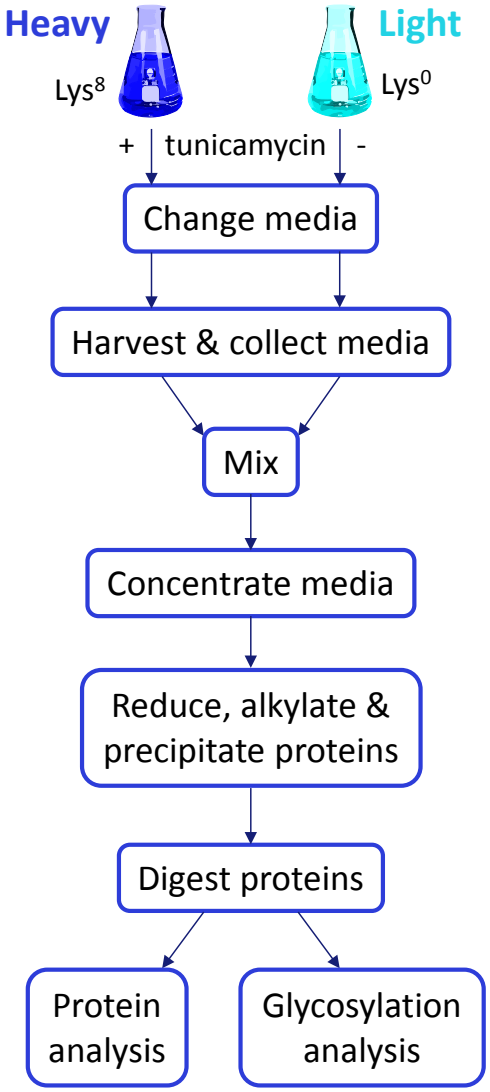

Figure 4.1. Experimental setup for each triplicate to study secreted proteins and glycoproteins in yeast cells by incorporating SILAC and N-glycosylation inhibition.

4.3.2 Example of secreted protein identification and quantification

An example of peptide identification and quantification are shown in Figure 4.2, including a full mass spectrum, tandem mass spectrum and extracted elution profiles for the heavy and light versions of the peptide YGSDGLSMTLAK* (* denotes heavy lysine). The light version of the peptide was identified from untreated cells and the heavy peptide was from treated cells; the full MS is displayed in Figure 4.2A. As shown in Figure 4.2B, the heavy peptide was very confidently identified with an XCorr of 3.65, and mass accuracy of 0.88 ppm. This peptide is from the protein YGR189C, probable glycosidase CRH1, which plays a role in cell wall architecture. Based on the elution profiles from light and heavy versions of the peptide, we were able to quantify the abundance change of this peptide with a ratio of 0.37 (Figure 4.2C). Because each of the elution profiles contains many scans, the ratio was highly accurate. This protein was quantified based on 43, 40 and 35 total peptides in each experiment, respectively, and the ratios are very consistent, i.e. 0.37, 0.39 and 0.37. All proteins identified and quantified in triplicate experiments are listed in the Supporting Information online at doi.org/10.1021/acs.jproteome.6b00953.

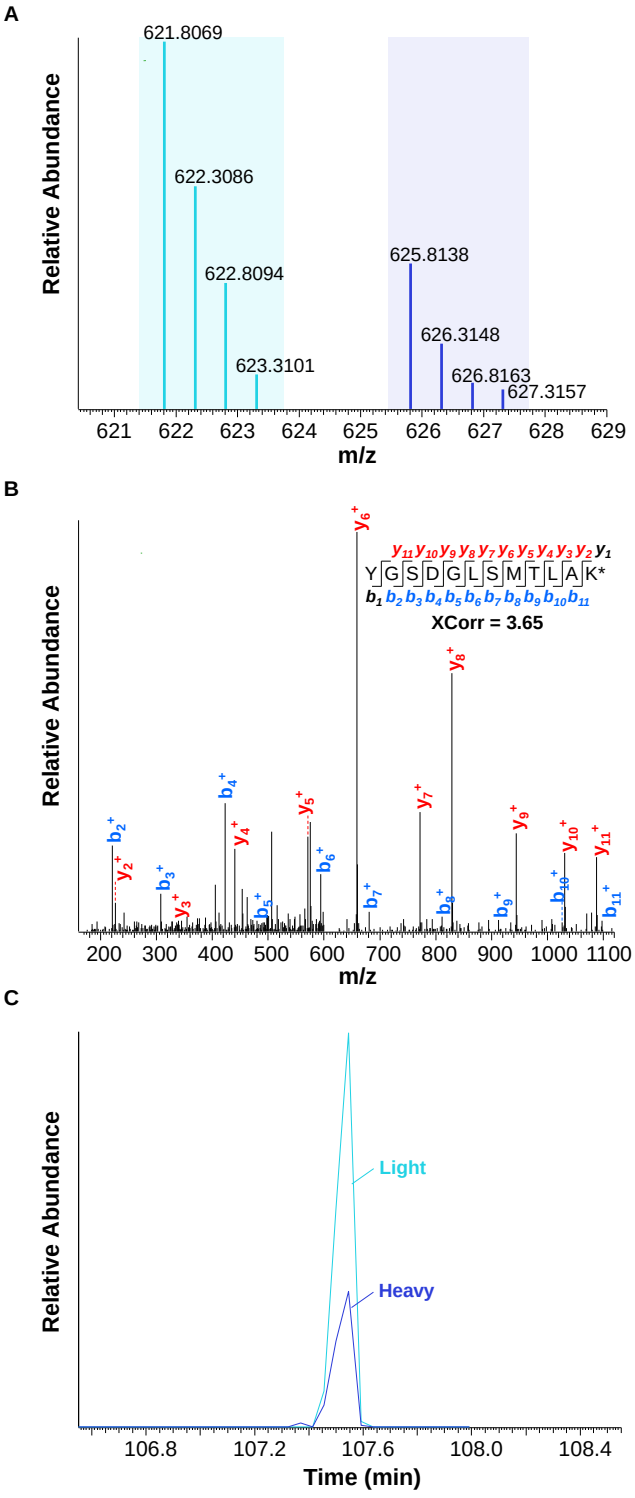

Figure 4.2. Example spectra for peptide identification and quantification. (A) Full mass spectrum showing heavy and light peptides, (B) tandem mass spectrum for peptide identification and (C) elution profiles of heavy and light versions of the peptide for quantification.

4.3.3 Secreted protein identification and quantification in biological triplicate experiments

Overall the number of proteins identified in each of the three triplicate experiments was very consistent: 198, 185 and 186 proteins, respectively. Figure 4.3A shows the overlap between proteins identified in each triplicate; 144 proteins were identified in all three experiments, and 180 were identified in at least two experiments. A total of 245 secreted proteins were identified.

Gene ontology analysis with DAVID was performed to determine which cellular components, molecular functions and biological processes corresponded to the 180 proteins identified in at least two triplicates and the results confirmed the presence of mainly secreted proteins (Figure 4.3). First, cellular component clustering revealed that the extracellular region was highly enriched with an extremely low *P* value of 5.90E-33, in addition to fungal-type cell wall (9.83E-36), external encapsulating structure (1.73E-35) and plasma membrane (1.38E-15). Interestingly, ribosomal proteins were also enriched. Previous reports indicate that some ribosomal proteins can be located on the cell surface or in the extracellular space, which is a possibility here.[66] Although they also could be due to cell death, the presence of many ribosomal proteins and their enrichment in the dataset compared to other intracellular proteins suggests that these proteins could be intentionally secreted. Molecular function and biological process analysis showed further cell wall and extracellular functions highly enriched in the dataset, including hydrolase activity, cell wall organization and glucanosyltransferase activity (Figure 4.3). Proteins with hydrolase activity (acting on glycosyl bonds) identified in at least two triplicates are listed in Table 4.1.

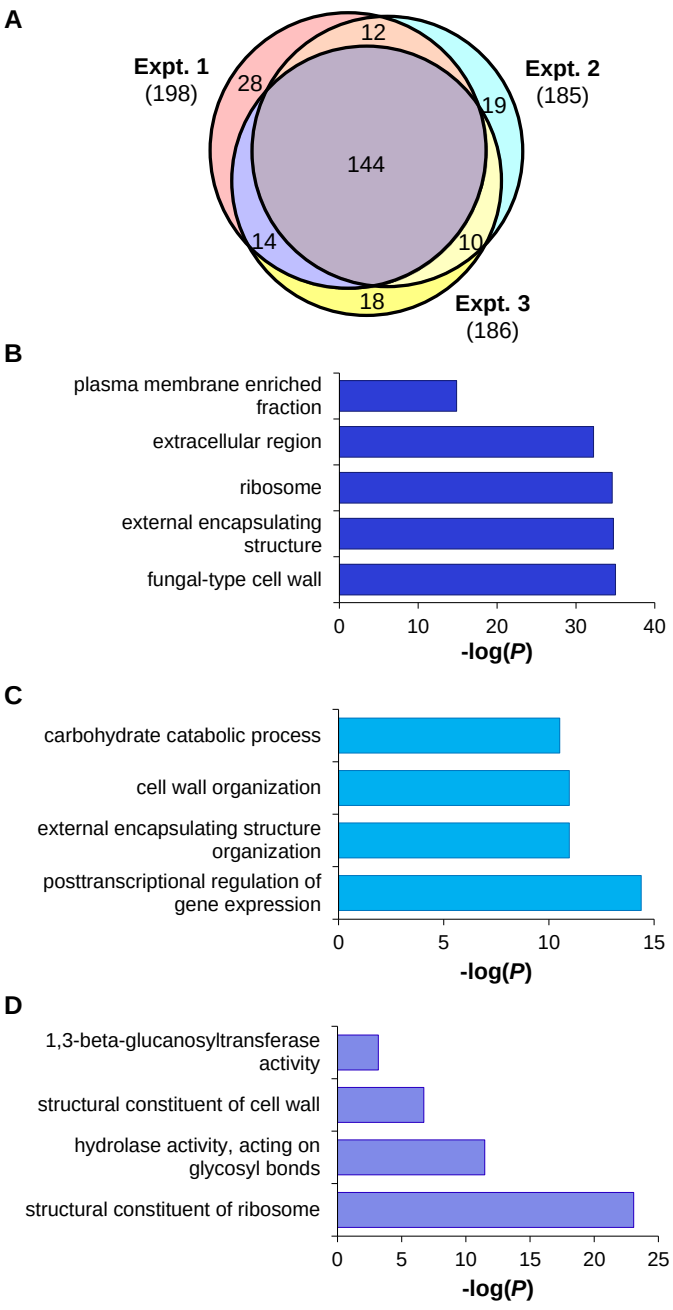

Figure 4.3. Proteins identified in each triplicate. (A) Overlap of proteins identified and clustering according to (B) cellular component, (C) biological process and (D) molecular function.

Table 4.1. Proteins identified in at least two triplicates clustered with the molecular function of hydrolase activity (acting on glycosyl bonds).

Gene symbol	Reference	Unique peptides	Total peptides	Protein ratio	Annotation
SGA1	YIL099W	2	8	0.02	Intracellular sporulation-specific glucoamylase involved in glycogen degradation
CRH1	YGR189C	9	118	0.37	Chitin transglycosylase that functions in the transfer of chitin to beta(1-6) and beta(1-3) glucans in the cell wall
SUN4	YNL066W	3	27	0.39	Cell wall protein related to glucanases, possibly involved in cell wall septation
SUC2	YIL162W	10	23	0.08	Invertase, sucrose hydrolyzing enzyme
ATH1	YPR026W	14	35	0.06	Acid trehalase required for utilization of extracellular trehalose
PGU1	YJR153W	4	12	1.45	Endo-polygalacturonase, pectolytic enzyme that hydrolyzes the alpha-1,4-glycosidic bonds in the rhamnogalacturonan chains in pectins
BGL2	YGR282C	8	60	0.64	Endo-beta-1,3-glucanase, major protein of the cell wall, involved in cell wall maintenance
EGT2	YNL327W	8	106	0.03	GPI-anchored cell wall endoglucanase required for proper cell separation after cytokinesis
SCW4	YGR279C	13	175	0.15	Cell wall protein with similarity to glucanases
EXG2	YDR261C	7	21	0.23	Exo-1,3-beta-glucanase, involved in cell wall beta-glucan assembly
EXG1	YLR300W	8	142	0.19	Major exo-1,3-beta-glucanase of the cell wall, involved in cell wall beta-glucan assembly
CTS1	YLR286C	3	30	0.34	Endochitinase, required for cell separation after mitosis
DSE4	YNR067C	13	60	0.46	Daughter cell-specific secreted protein with similarity to glucanases
SCW10	YMR305C	10	80	0.89	Cell wall protein with similarity to glucanases
UTR2	YEL040W	3	37	0.31	Chitin transglycosylase that functions in the transfer of chitin to beta(1-6) and beta(1-3) glucans in the cell wall
SCW11	YGL028C	5	192	0.25	Cell wall protein with similarity to glucanases
DSE2	YHR143W	1	7	0.48	Daughter cell-specific secreted protein with similarity to glucanases

A total of 192, 179 and 179 proteins were quantified in each of the three triplicate experiments, respectively. Most of them were down-regulated, *i.e.* 151, 145 and 134 proteins, in tunicamycin-treated cells labeled with heavy lysine. Protein ratios quantified in the three experiments are highly reproducible, as demonstrated in Figure 4.4A through comparison between the first and second experiments (R=0.92) and the first and third experiments (R=0.93). Deviations could be due to variations in protein abundances among cells in biological triplicate experiments. In addition, sample preparation and MS measurements could also result in experimental errors.

A total of 239 proteins were quantified in the triplicate experiments. 174 proteins were quantified in at least two experiments, among which 135 were down-regulated and 21 were up-regulated; the protein median ratio distributions are shown in Figure 4.4B. For down-regulated proteins, their secretion is very likely regulated by glycosylation. After protein glycosylation is inhibited by tunicamycin, proteins will not be folded properly and as a result will be trapped inside cells. For proteins with increased abundances after the treatment, it is clear that their secretion is not dependent on protein glycosylation. Their expression was up-regulated due to cellular responses to the tunicamycin treatment. For example, HSP12 was up-regulated by 3.4 fold in the secretome. This is a plasma membrane localized protein that protects membranes from desiccation; and its expression is induced by external stimulus including heat shock, oxidative stress and osmostress. In our previous research, this protein was up-regulated by 3.1 fold in cells treated by tunicamycin for the whole cell lysate analysis.[47] This is further supported by gene ontology clustering according to biological process. As shown in Figure 4.5, among 21 up-regulated proteins, eight proteins were related to response to

oxidative stress, which is the most highly enriched group with a *P* value of 8.0E-9. Biological processes corresponding to cell redox homeostasis, response to temperature stimulus and homeostatic process, were also highly enriched in up-regulated proteins.

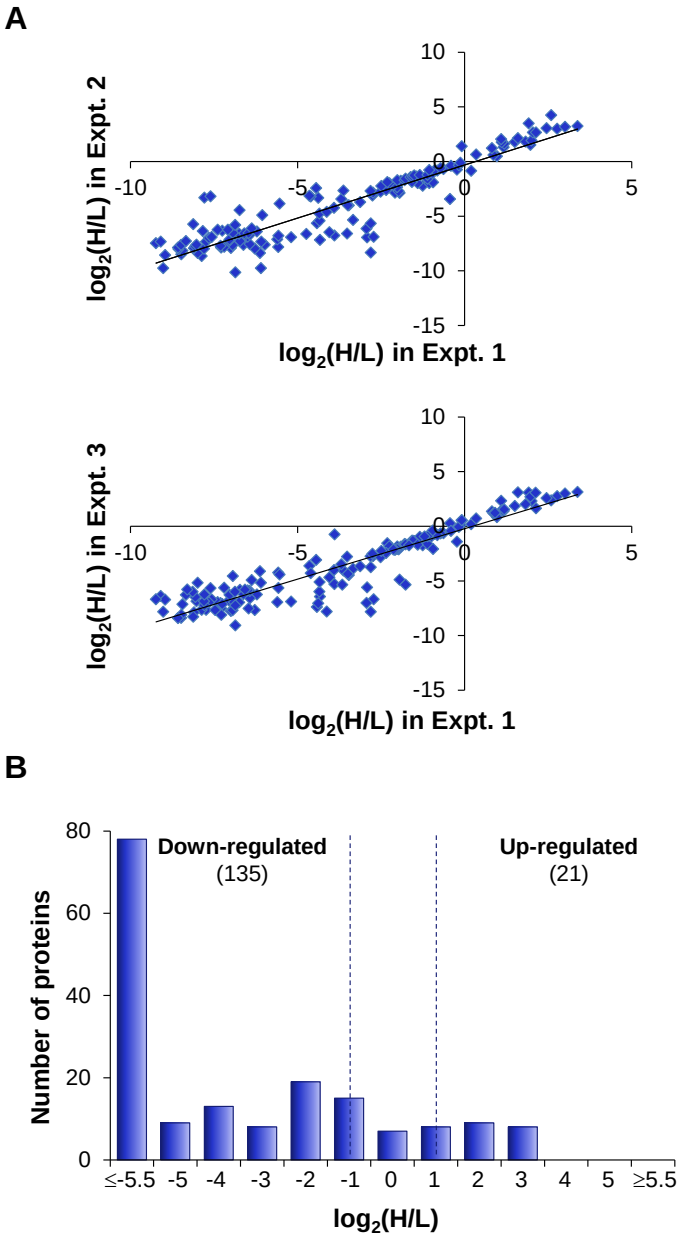

Figure 4.4. Secreted proteins quantified in all three triplicates. (A) Comparison of quantified secreted proteins between the first and second experiments (top) and the first and third experiments (bottom), and (B) the median ratio distribution of secreted proteins quantified in at least two experiments.

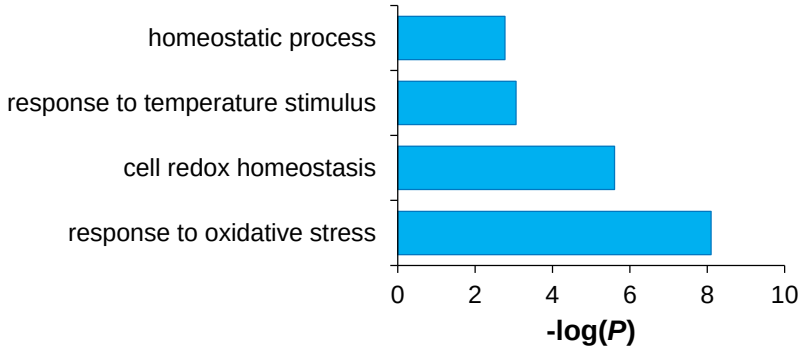

Figure 4.5. Clustering of 21 up-regulated proteins quantified in at least two triplicates according to biological process.

This study provides the first systematic identification of secreted proteins in *Saccharomyces cerevisiae*. A total of 245 secreted proteins were identified; 144 proteins were found in all three experiments and 180 proteins were identified in at least two experiments. After cells were treated with tunicamycin for two hours, secreted protein quantification results demonstrated that many proteins were down-regulated. The down-regulation of these proteins strongly suggests they are secreted through the classical pathway and their secretion is very likely regulated by protein N-glycosylation.

4.3.4 Site specific analysis of protein glycosylation in the secretome

The enrichment of glycopeptides is essential for their MS-based identification and quantification, and the common tag generated through treatment with PNGase F in heavy-oxygen water enabled us to confidently identify glycopeptides and localize the glycosylation sites. An example of glycopeptide identification (YSRCDTLVGN#LTIGGGLK, # represents glycosylation site) is shown in Figure 4.6.

This glycopeptide was confidently identified with an XCorr value of 5.58 and mass accuracy of 0.98 ppm. Based on the fragments, the glycosylation site was well localized at N57 with the well-known motif of NXS/T (X is any amino acid residue except proline). This glycopeptide is from the cell wall mannoprotein PST1 (YDR055W), which is a glycosylphosphatidylinositol (GPI)-anchored plasma membrane protein. Based on the information on UniProt, this protein is highly N-glycosylated and the site S419 is covalently attached to GPI. This glycopeptide was also quantified and down-regulated in all three glycosylation triplicates (0.13, 0.14 and 0.14, respectively). All glycosylation sites and glycopeptides quantified in triplicate experiments are listed in the Supporting Information online at doi.org/10.1021/acs.jproteome.6b00953.

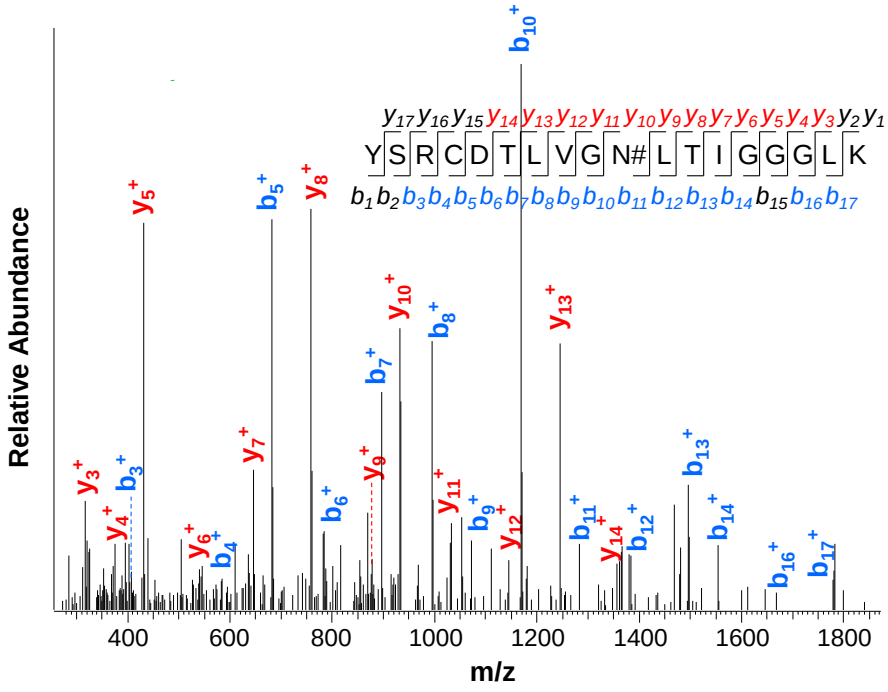

Figure 4.6. Tandem mass spectrum for the identification of the glycopeptide YSRCDTLVGN#LTIGGGLK (# denotes glycosylation site).

A total of 110 unique glycosylation sites were identified in three glycosylation triplicate experiments, with 92, 78 and 83 sites in each, respectively, and their overlap is shown in Figure 4.7A. There were 60 glycosylation sites identified in all three triplicates, and 20 sites were localized to several domains, including thioredoxin, PLA2c, and peptidase a1 domains. Overall, the majority of glycosylation sites identified were not located within any domain (82%), as shown in Figure 4.7B (protein domain information is from UniProt).

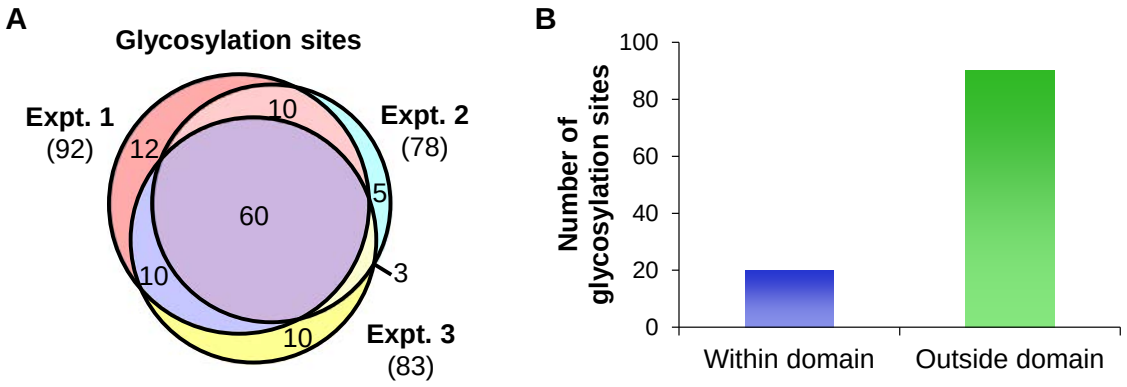

Figure 4.7. (A) Comparison of glycosylation site identification in triplicate experiments, and (B) the numbers of glycosylation sites (in at least two experiments) located within a domain or outside of any domain.

Similarly, the numbers of glycopeptides and glycoproteins quantified were consistent in the glycosylation triplicate experiments: 89, 85, and 84 unique glycopeptides and 38, 36 and 38 glycoproteins, respectively. The overlap between glycopeptides quantified in each triplicate is shown in Figure 4.8. A total of 119 unique glycopeptides were quantified, and 83 were quantified in at least two; their ratio

distributions are shown in Figure 4.9A. Overall 47 glycoproteins were identified and quantified in the three triplicate experiments, 36 were quantified in at least two, and their protein ratio distribution is shown in Figure 4.9B. A total of 29 common glycoproteins were identified in all triplicate glycosylation experiments, with the majority of protein ratios down-regulated by more than two fold (27 down-regulated). The abundances of the other two proteins also decreased with ratios of 0.51 and 0.71; likely the secretion of these two proteins was also regulated by their glycosylation, but they may be relatively stable during treatment.

Theoretically all glycopeptides and glycoproteins should be down-regulated after protein N-glycosylation inhibition by tunicamycin. However, several were not down-regulated. The possible reasons include: the inhibition of protein N-glycosylation may not be complete; the parent protein abundance may be increased; some glycoproteins could be very stable, and their protein synthesis and degradation may not be rapid enough in terms of the time scale of the treatment.

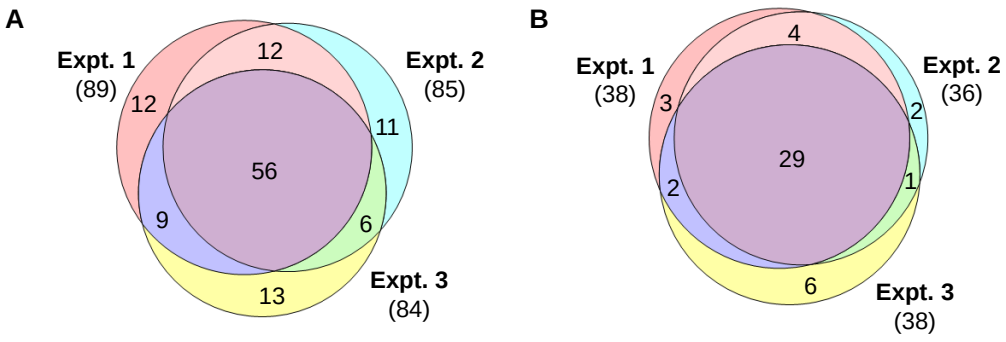

Figure 4.8. Overlap of (A) glycopeptides and (B) glycoproteins in all three glycosylation experiments.

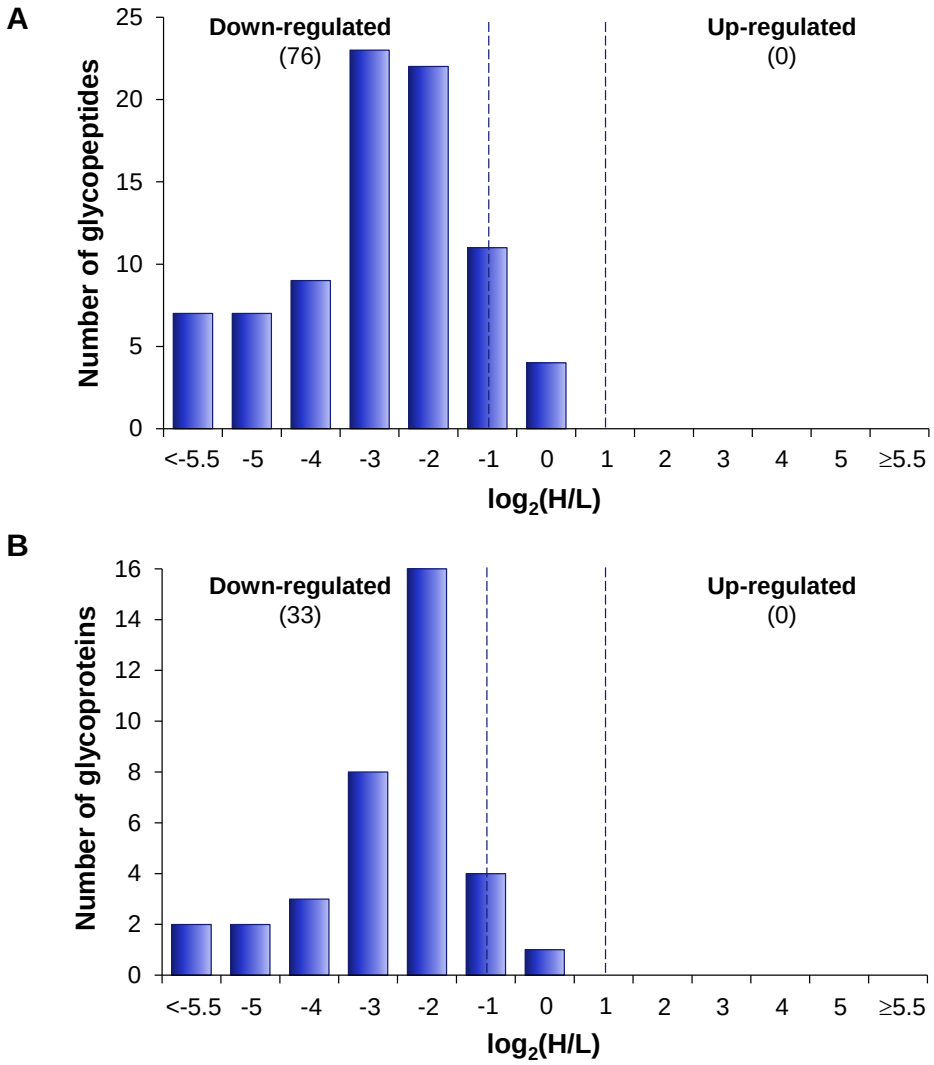

Figure 4.9. Ratio distribution of (A) glycopeptides and (B) glycoproteins quantified in at least two triplicates.

4.3.5 Clustering of glycoproteins

The gene ontology clustering of the 36 overlapped proteins (using glycoproteins identified in at least two experiments) indicates that the most highly enriched cellular

components were all related to cell membrane or extracellular categories, and vacuolar components (Figure 4.10), which also are involved in secretion. Protein clustering based on molecular function and biological process had similar results, including cell wall organization, carbohydrate metabolic processes and hydrolase activities, which are all expected to be involved in protein secretion or extracellular activities.

4.3.6 GPI-anchored proteins and glycoproteins in the secretome

GPI-modified proteins are typically anchored into the cell plasma membrane through the lipid component of the modified GPI. This group of proteins is normally located on the outside of the cell surface, and if the modified group gets cleaved or hydrolyzed, the protein will be released into the secretome. In this work, we identified a total of 14 GPI modified glycoproteins and 63 unique glycopeptides from these GPI-anchored proteins. Several glycopeptides from GPI-anchored proteins are listed in Table 4.2, along with their corresponding protein ratios. For example, GAS1, 1,3-beta-glucanosyltransferase, is involved in cell wall biosynthesis and morphogenesis. The site N528 has been reported to be modified with GPI.[67] In our experiments, we identified 6 unique glycopeptides from this protein, and the glycopeptide with the site N95 is listed in Table 4.2. This glycopeptide was down-regulated with a median ratio of 0.22. The protein was also down-regulated in the secretome of the treated cells. All glycopeptides from GPI-anchored proteins are listed in the Supporting Information online at doi.org/10.1021/acs.jproteome.6b00953.

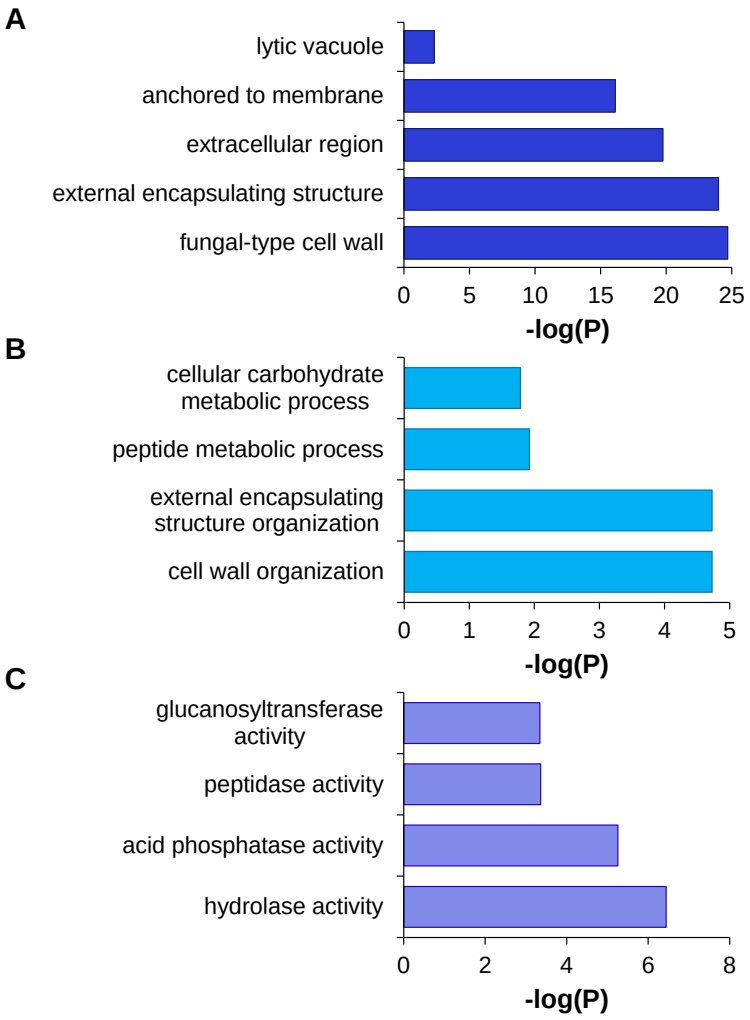

Figure 4.10. Clustering of glycoproteins in at least two triplicates according to (A) cellular component, (B) biological process and (C) molecular function.

Table 4.2. Several examples of glycopeptides quantified in at least two glycosylation triplicates from GPI-anchored proteins (*P* = 7.81E-17) (# denotes glycosylation site, and * denotes heavy lysine).

Gene Symbol	Peptide	Site	ModScore	Peptide ratio	Standard deviation	Protein ratio	Annotation
ECM33	VQTVGGAIEVTGN#FSTLDLSSLK	304	1000	0.13	0.005	0.14	Cell wall protein ECM33
	SVRGGANFDSSSSN#FSCNALK	328	47.1	0.33	0.043		
PST1	YSRCDTLVGN#LTIGGGLK	57	1000	0.14	0.008	0.14	Cell wall mannoprotein PST1
	IGGLDN#LTTIGGTLEVVGN#FTSLNLDSLK	292	120.5	0.14	0.012		
		305	34.1				
CRH1	FHN#YTLDWAMDK	177	1000	0.13	0.007	0.13	Probable glycosidase CRH1
YPS3	RN#ITLTTTK	275	1000	0.04	0.001	0.03	Aspartic proteinase yapsin-3
	SLN#ASYSK	309	1000	0.03	0.005		
GAS1	LNTNVIRVYAIN#TTLDHSECMK*	95	63.4	0.22	0.011	0.33	1,3-beta-glucanosyltransferase GAS1
PLB3	YLGTN#VSNGVPLERGK	313	8.0	0.07	0.008	0.07	Lysophospho-lipase 3
	NYCWN#GTLDTTPLPDVEK	588	42.1	0.03	0.103		

Another example is PLB3, phospholipase B (lysophospholipase), which is involved in phospholipid metabolism, hydrolyzes phosphatidylinositol and phosphatidylserine, and displays transacylase activity *in vitro*. We identified 5 unique glycopeptides, and two glycopeptides identified in all three triplicates are listed in Table 2. Their ratios are 0.07 and 0.03, respectively, which further demonstrates the site specificity of this research. The majority of these glycopeptides were highly down-regulated (all except one), and similarly the protein was also down-regulated in the secretome by 14.9 fold. In our previous experiment, this protein was up-regulated by 8.0 fold inside of tunicamycin-treated cells.[47] This is extremely consistent with the experimental design because the inhibition of N-glycosylation prevents the protein from being secreted, and therefore it accumulates inside cells. Examples such as this provide solid evidence that this protein's secretion is regulated by its N-glycosylation.

4.3.7 Comparison of secreted proteins and glycoproteins

The glycoproteins quantified in at least two glycosylation triplicates were compared to the proteins quantified in at least two protein triplicates, and 27 proteins were overlapped (Figure 4.11A). The nine glycoproteins not identified in the triplicate protein experiments could be due to their low abundance and resultant lack of identification without any enrichment. Further investigation of these 27 proteins revealed that 25 were previously reported to be secreted (located in the extracellular matrix, plasma membrane, or containing a signal peptide), and the other two proteins were localized to the vacuole, indicating that they may also be secreted. All proteins except one were consistently down-regulated. The protein PRY3, whose function is unknown, was up-regulated at protein level with a ratio of 2.76, but the glycopeptide

YN#YSNPGFSESTGHFTQVVWK from this protein was down-regulated with a ratio of 0.31. After protein glycosylation was inhibited, the glycopeptide was expected to be down-regulated. However, the increased abundance in the secretome means that this protein's secretion is independent from its glycosylation. Clustering of 27 proteins present in all protein and glycosylation triplicates based on cellular component showed the extracellular region and cell wall to be highly enriched (*P* values of 2.79E-21 and 9.51E-23, respectively).

The majority of quantified proteins in at least two experiments, 135 of 174 total (77.6%), were down-regulated in the secretome of cells treated with tunicamycin. Most likely the secretion of these proteins was regulated by their glycosylation, since without N-glycosylation, they may not be properly folded and secreted. Among all down-regulated proteins, 27 were also identified as glycosylated in the secretome, and nearly all of them (26 proteins) were down-regulated. This is another piece of solid evidence that the secretion of these 26 proteins were dependent on their N-glycosylation.

4.3.8 Abundance distribution of identified secreted proteins

Abundances of the 180 proteins identified in at least two protein triplicates were investigated to determine the depth of proteins quantified in this experiment. Figure 4.11B shows the distribution of these proteins, and their abundances are from the literature.[68] Although some proteins identified here are high abundance proteins (greater than 10,000 copies per cell), we also identified many low-abundance proteins. For example, 40 out of 180 proteins (22%) were not visualized and 8 proteins (4%) were not quantified in the previous report. From the distribution, the abundances of proteins

identified span a wide range. Due to the low abundance of many secreted proteins, if more starting materials are used, more proteins with low abundance should be identified.

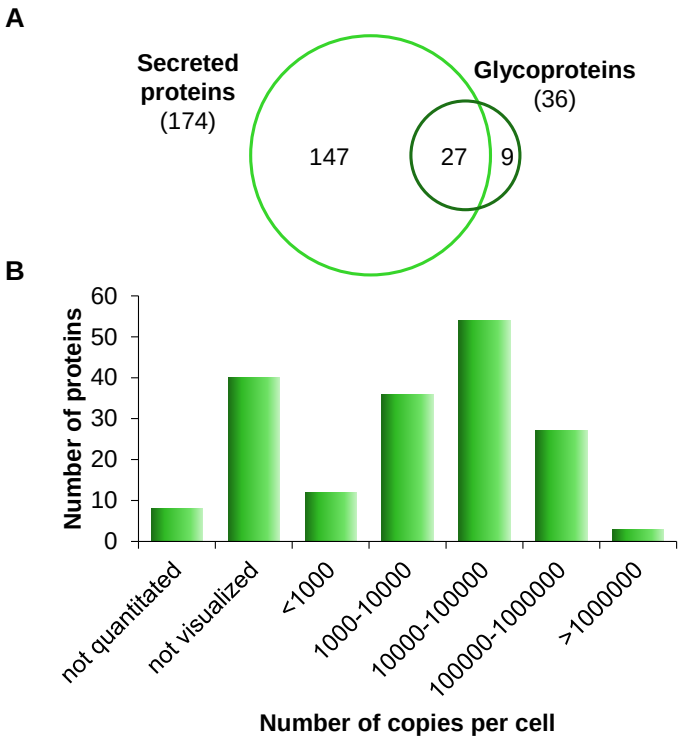

Figure 4.11. (A) Overlap between secreted proteins and glycoproteins quantified in at least two experiments. (B) Abundance distribution of 180 proteins identified in at least two protein triplicates according to protein abundance in the literature.[68]

4.4 Conclusions

Protein secretion is essential for cells to interact with the surrounding environment and other cells, and secreted proteins are a promising and non-invasive source of biomarkers for disease detection. Systematic investigation of protein secretion has

profound implications in biological and biomedical research. By incorporating SILAC labeling, tunicamycin treatment and MS-based proteomics, secreted proteins in yeast have been comprehensively identified and quantified, and their glycosylation has been investigated for the first time. A total of 245 proteins were identified in triplicate experiments, with 144 common proteins identified in each of the three experiments. Protein clustering results indicated that proteins in the extracellular space were highly enriched. After cells were treated with a potent protein N-glycosylation inhibitor, tunicamycin, the majority of quantified secreted proteins were down-regulated, which strongly suggests that the secretion of these proteins was regulated by their N-glycosylation, while the secretion of some proteins with minimal abundance changes was not related to glycosylation. In addition, we also systematically identified and quantified protein N-glycosylation in the secretome. Totally 110 glycosylation sites were located on 47 secreted proteins. Nearly all quantified glycoproteins were down-regulated, which is consistent with the inhibition of protein N-glycosylation. By using yeast as a model eukaryotic system, we performed the first systematic study of protein secretion and its regulation by N-glycosylation. These experimental results provide valuable information about protein secretion and advance our understanding of protein secretion and the functions of protein N-glycosylation.

4.6 References

1. Lee, M. C. S.; Miller, E. A.; Goldberg, J.; Orci, L.; Schekman, R., Bi-directional protein transport between the ER and Golgi. *Annu Rev Cell Dev Bi* **2004,** *20*, 87-123.

2. Makridakis, M.; Vlahou, A., Secretome proteomics for discovery of cancer biomarkers. *J Proteomics* **2010,** *73* (12), 2291-2305.

3. Pavlou, M. P.; Diamandis, E. P., The cancer cell secretome: A good source for discovering biomarkers? *J Proteomics* **2010,** *73* (10), 1896-1906.

4. Nickel, W.; Seedorf, M., Unconventional Mechanisms of Protein Transport to the Cell Surface of Eukaryotic Cells. *Annu Rev Cell Dev Bi* **2008,** *24*, 287-308.

5. Nombela, C.; Gil, C.; Chaffin, W. L., Non-conventional protein secretion in yeast. *Trends Microbiol.* **2006,** *14* (1), 15-21.

6. Giardina, B. J.; Stanley, B. A.; Chiang, H. L., Glucose induces rapid changes in the secretome of Saccharomyces cerevisiae. *Proteome Sci.* **2014,** *12*, 21.

7. Pandey, A.; Podtelejnikov, A. V.; Blagoev, B.; Bustelo, X. R.; Mann, M.; Lodish, H. F., Analysis of receptor signaling pathways by mass spectrometry: Identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *P Natl Acad Sci USA* **2000,** *97* (1), 179-184.

8. Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W., Etzler, M.E., *Essentials of Glycobiology*. 2 ed.; Cold Spring Harbor Laboratory Press: New York, 2009.

9. Haltiwanger, R. S.; Lowe, J. B., Role of glycosylation indevelopment. *Annu. Rev. Biochem.* **2004,** *73*, 491-537.

10. Braakman, I.; Bulleid, N. J., Protein Folding and Modification in the Mammalian Endoplasmic Reticulum. In *Annual Review of Biochemistry, Vol 80*, Kornberg, R. D.; Raetz, C. R. H.; Rothman, J. E.; Thorner, J. W., Eds. Annual Reviews: Palo Alto, 2011; Vol. 80, pp 71-99.

11. Schwarz, F.; Aebi, M., Mechanisms and principles of N-linked protein glycosylation. *Curr. Opin. Struct. Biol.* **2011,** *21* (5), 576-582.

12. Liang, Y. X.; Eng, W. S.; Colquhoun, D. R.; Dinglasan, R. R.; Graham, D. R.; Mahal, L. K., Complex N-Linked Glycans Serve as a Determinant for Exosome/Microvesicle Cargo Recruitment. *J. Biol. Chem.* **2014,** *289* (47), 32526–32537.

13. Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H., SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **2011,** *8* (10), 785-786.

14. Tjalsma, H.; Bolhuis, A.; Jongbloed, J. D. H.; Bron, S.; van Dijl, J. M., Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiology and Molecular Biology Reviews* **2000,** *64* (3), 515-547.

15. Marimuthu, A.; Chavan, S.; Sathe, G.; Sahasrabuddhe, N. A.; Srikanth, S. M.; Renuse, S.; Ahmad, S.; Radhakrishnan, A.; Barbhuiya, M. A.; Kumar, R. V.; Harsha, H. C.; Sidransky, D.; Califano, J.; Pandey, A.; Chatterjee, A., Identification of head and neck squamous cell carcinoma biomarker candidates through proteomic analysis of cancer cell secretome. *BBA-Proteins Proteomics* **2013,** *1834* (11), 2308-2316.

16. Boersema, P. J.; Geiger, T.; Wisniewski, J. R.; Mann, M., Quantification of the N-glycosylated Secretome by Super-SILAC During Breast Cancer Progression and in Human Blood Samples. *Mol. Cell. Proteomics* **2013,** *12* (1), 158-171.

17. Brown, K. J.; Formolo, C. A.; Seol, H.; Marathi, R. L.; Duguez, S.; An, E.; Pillai, D.; Nazarian, J.; Rood, B. R.; Hathout, Y., Advances in the proteomic investigation of the cell secretome. *Expert Rev. Proteomics* **2012,** *9* (3), 337-345.

18. Yin, X. K.; Bern, M.; Xing, Q. R.; Ho, J.; Viner, R.; Mayr, M., Glycoproteomic Analysis of the Secretome of Human Endothelial Cells. *Mol. Cell. Proteomics* **2013,** *12* (4), 956-978.

19. Hasan, N. M.; Adams, G. E.; Joiner, M. C., Effect of serum starvation on expression and phosphorylation of PKC-alpha and p53 in V79 cells: Implications for cell death. *Int. J. Cancer* **1999,** *80* (3), 400-405.

20. Shin, J. S.; Hong, S. W.; Lee, S. L. O.; Kim, T. H.; Park, I. C.; An, S. K.; Lee, W. K.; Lim, J. S.; Kim, K. I.; Yang, Y.; Lee, S. S.; Jin, D. H.; Lee, M. S., Serum starvation induces G1 arrest through suppression of Skp2-CDK2 and CDK4 in SK-OV-3 cells. *Int. J. Oncol.* **2008,** *32* (2), 435-439.

21. Levin, V. A.; Panchabhai, S. C.; Shen, L.; Kornblau, S. M.; Qiu, Y. H.; Baggerly, K. A., Different Changes in Protein and Phosphoprotein Levels Result from Serum Starvation of High-Grade Glioma and Adenocarcinoma Cell Lines. *J. Proteome Res.* **2010,** *9* (1), 179-191.

22. Washburn, M. P.; Wolters, D.; Yates, J. R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** *19* (3), 242-247.

23. Witze, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G., Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* **2007,** *4* (10), 798-806.

24. Sun, L. L.; Dubiak, K. M.; Peuchen, E. H.; Zhang, Z. B.; Zhu, G. J.; Huber, P. W.; Dovichi, N. J., Single Cell Proteomics Using Frog (Xenopus laevis) Blastomeres Isolated from Early Stage Embryos, Which Form a Geometric Progression in Protein Content. *Anal. Chem.* **2016,** *88* (13), 6653-6657.

25. Pankow, S.; Bamberger, C.; Calzolari, D.; Martinez-Bartolome, S.; Lavallee-Adam, M.; Balch, W. E.; Yates, J. R., Delta F508 CFTR interactome remodelling promotes rescue of cystic fibrosis. *Nature* **2015,** *528* (7583), 510-516.

26. Ma, J. F.; Banerjee, P.; Whelan, S. A.; Liu, T.; Wei, A. C.; Ramirez-Correa, G.; McComb, M. E.; Costello, C. E.; O'Rourke, B.; Murphy, A.; Hart, G. W., Comparative Proteomics Reveals Dysregulated Mitochondrial O-GlcNAcylation in Diabetic Hearts. *J. Proteome Res.* **2016,** *15* (7), 2254-2264.

27. Ji, Y. H.; Leymarie, N.; Haeussler, D. J.; Bachschmid, M. M.; Costello, C. E.; Lin, C., Direct Detection of S-Palmitoylation by Mass Spectrometry. *Anal. Chem.* **2013,** *85* (24), 11952-11959.

28. Hong, Q. T.; Ruhaak, L. R.; Stroble, C.; Parker, E.; Huang, J. C.; Maverakis, E.; Lebrilla, C. B., A Method for Comprehensive Glycosite-Mapping and Direct Quantitation of Serum Glycoproteins. *J. Proteome Res.* **2015,** *14* (12), 5179-5192.

29. Brodbelt, J. S., Ion Activation Methods for Peptides and Proteins. *Anal. Chem.* **2016,** *88* (1), 30-51.

30. Fort, K. L.; Dyachenko, A.; Potel, C. M.; Corradini, E.; Marino, F.; Barendregt, A.; Makarov, A. A.; Scheltema, R. A.; Heck, A. J. R., Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics. *Anal. Chem.* **2016,** *88* (4), 2303-2310.

31. Yu, C.; Yang, Y. Y.; Wang, X. R.; Guan, S. H.; Fang, L.; Liu, F.; Walters, K. J.; Kaiser, P.; Huang, L., Characterization of Dynamic UbR-Proteasome Subcomplexes by In vivo Cross-linking (X) Assisted Bimolecular Tandem Affinity Purification (XBAP) and Label-free Quantitation. *Mol. Cell. Proteomics* **2016,** *15* (7), 2279-2292.

32. Thaysen-Andersen, M.; Mysling, S.; Hojrup, P., Site-Specific Glycoprofiling of N-Linked Glycopeptides Using MALDI-TOF MS: Strong Correlation between Signal Strength and Glycoform Quantities. *Anal. Chem.* **2009,** *81* (10), 3933-3943.

33. Ludwig, K. R.; Dahl, R.; Hummon, A. B., Evaluation of the mirn23a Cluster through an iTRAQ-based Quantitative Proteomic Approach. *J. Proteome Res.* **2016,** *15* (5), 1497-1505.

34. Mukherjee, P.; Mani, S., Methodologies to decipher the cell secretome. *BBA-Proteins Proteomics* **2013,** *1834* (11), 2226-2232.

35. Sorgo, A. G.; Heilmann, C. J.; Dekker, H. L.; Brul, S.; de Koster, C. G.; Klis, F. M., Mass spectrometric analysis of the secretome of *Candida albicans*. *Yeast* **2010,** *27* (8), 661-672.

36. Sun, S. S.; Shah, P.; Eshghi, S. T.; Yang, W. M.; Trikannad, N.; Yang, S.; Chen, L. J.; Aiyetan, P.; Hoti, N.; Zhang, Z.; Chan, D. W.; Zhang, H., Comprehensive analysis

of protein glycosylation by solid-phase extraction of N-linked glycans and glycosite-containing peptides. *Nat. Biotechnol.* **2016,** *34* (1), 84-88.

37. Wang, X. S.; Yuan, Z. F.; Fan, J.; Karch, K. R.; Ball, L. E.; Denu, J. M.; Garcia, B. A., A Novel Quantitative Mass Spectrometry Platform for Determining Protein O-GlcNAcylation Dynamics. *Mol. Cell. Proteomics* **2016,** *15* (7), 2462-2475.

38. Colak, G.; Pougovkina, O.; Dai, L. Z.; Tan, M. J.; te Brinke, H.; Huang, H.; Cheng, Z. Y.; Park, J.; Wan, X. L.; Liu, X. J.; Yue, W. W.; Wanders, R. J. A.; Locasale, J. W.; Lombard, D. B.; de Boer, V. C. J.; Zhao, Y. M., Proteomic and Biochemical Studies of Lysine Malonylation Suggest Its Malonic Aciduria-associated Regulatory Role in Mitochondrial Function and Fatty Acid Oxidation. *Mol. Cell. Proteomics* **2015,** *14* (11), 3056-3071.

39. Wu, R. H.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Methods* **2011,** *8* (8), 677-683.

40. Ramya, T. N. C.; Weerapana, E.; Cravatt, B. F.; Paulson, J. C., Glycoproteomics enabled by tagging sialic acid- or galactose-terminated glycans. *Glycobiology* **2013,** *23* (2), 211-221.

41. Woo, C. M.; Iavarone, A. T.; Spiciarich, D. R.; Palaniappan, K. K.; Bertozzi, C. R., Isotope-targeted glycoproteomics (IsoTaG): a mass-independent platform for intact N- and O-glycopeptide discovery and analysis. *Nat Methods* **2015,** *12* (6), 561-567.

42. Martin, B. R.; Cravatt, B. F., Large-scale profiling of protein palmitoylation in mammalian cells. *Nat Methods* **2009,** *6* (2), 135-138.

43. Mertins, P.; Qiao, J. W.; Patel, J.; Udeshi, N. D.; Clauser, K. R.; Mani, D. R.; Burgess, M. W.; Gillette, M. A.; Jaffe, J. D.; Carr, S. A., Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods* **2013,** *10* (7), 634-637.

44. Zielinska, D. F.; Gnad, F.; Schropp, K.; Wisniewski, J. R.; Mann, M., Mapping N-Glycosylation Sites across Seven Evolutionarily Distant Species Reveals a Divergent Substrate Proteome Despite a Common Core Machinery. *Mol. Cell* **2012,** *46* (4), 542-548.

45. Alvarez-Manilla, G.; Warren, N. L.; Atwood, J.; Orlando, R.; Dalton, S.; Pierce, M., Glycoproteomic Analysis of Embryonic Stem Cells: Identification of Potential Glycobiomarkers Using Lectin Affinity Chromatography of Glycopeptides. *J. Proteome Res.* **2010,** *9* (5), 2062-2075.

46. Wei, X.; Dulberger, C.; Li, L. J., Characterization of Murine Brain Membrane Glycoproteins by Detergent Assisted Lectin Affinity Chromatography. *Anal. Chem.* **2010,** *82* (15), 6329-6333.

47. Xiao, H. P.; Smeekens, J. M.; Wu, R. H., Quantification of tunicamycin-induced protein expression and N-glycosylation changes in yeast. *Analyst* **2016,** *141* (12), 3737-3745.

48. Spiro, R. G., Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **2002,** *12* (4), 43r-56r.

49. Hu, H.; Khatri, K.; Klein, J.; Leymarie, N.; Zaia, J., A review of methods for interpretation of glycopeptide tandem mass spectral data. *Glycoconjugate J.* **2016,** *33* (3), 285-296.

50. Song, E.; Mayampurath, A.; Yu, C. Y.; Tang, H. X.; Mechref, Y., Glycoproteomics: Identifying the Glycosylation of Prostate Specific Antigen at Normal and High Isoelectric Points by LC-MS/MS. *J. Proteome Res.* **2014,** *13* (12), 5570-5580.

51. Zhu, Z. K.; Desaire, H., Carbohydrates on Proteins: Site-Specific Glycosylation Analysis by Mass Spectrometry. In *Annual Review of Analytical Chemistry, Vol 8*, Cooks, R. G.; Pemberton, J. E., Eds. Annual Reviews: Palo Alto, 2015; Vol. 8, pp 463-483.

52. Plomp, R.; Bondt, A.; de Haan, N.; Rombouts, Y.; Wuhrer, M., Recent Advances in Clinical Glycoproteomics of Immunoglobulins (Igs). *Mol. Cell. Proteomics* **2016,** *15* (7), 2217-2228.

53. Song, X. Z.; Ju, H.; Lasanajak, Y.; Kudelka, M. R.; Smith, D. F.; Cummings, R. D., Oxidative release of natural glycans for functional glycomics. *Nat Methods* **2016,** *13* (6), 528-534.

54. Walker, S. H.; Taylor, A. D.; Muddiman, D. C., Individuality Normalization when Labeling with Isotopic Glycan Hydrazide Tags (INLIGHT): A Novel Glycan-Relative Quantification Strategy. *J. Am. Soc. Mass Spectrom.* **2013,** *24* (9), 1376-1384.

55. Stockmann, H.; Duke, R. M.; Martin, S. M.; Rudd, P. M., Ultrahigh Throughput, Ultrafiltration-Based N-Glycomics Platform for Ultraperformance Liquid Chromatography (ULTRA(3)). *Anal. Chem.* **2015,** *87* (16), 8316-8322.

56. Kailemia, M. J.; Ruhaak, L. R.; Lebrilla, C. B.; Amster, I. J., Oligosaccharide Analysis by Mass Spectrometry: A Review of Recent Developments. *Anal. Chem.* **2014,** *86* (1), 196-212.

57. Sun, X. J.; Lin, L.; Liu, X. Y.; Zhang, F. M.; Chi, L. L.; Xia, Q. W.; Linhardt, R. J., Capillary Electrophoresis-Mass Spectrometry for the Analysis of Heparin Oligosaccharides and Low Molecular Weight Heparin. *Anal. Chem.* **2016,** *88* (3), 1937-1943.

58. Wang, L.; Aryal, U. K.; Dai, Z. Y.; Mason, A. C.; Monroe, M. E.; Tian, Z. X.; Zhou, J. Y.; Su, D.; Weitz, K. K.; Liu, T.; Camp, D. G.; Smith, R. D.; Baker, S. E.; Qian, W. J., Mapping N-Linked Glycosylation Sites in the Secretome and Whole Cells of

Aspergillus niger Using Hydrazide Chemistry and Mass Spectrometry. *J. Proteome Res.* **2012,** *11* (1), 143-156.

59. Lee, L. Y.; Thaysen-Andersen, M.; Baker, M. S.; Packer, N. H.; Hancock, W. S.; Fanayan, S., Comprehensive N-Glycome Profiling of Cultured Human Epithelial Breast Cells Identifies Unique Secretome N-Glycosylation Signatures Enabling Tumorigenic Subtype Classification. *J. Proteome Res.* **2014,** *13* (11), 4783-4795.

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

61. Wessel, D.; Flugge, U. I., A Method for the Quantitative Recovery of Protein in Dilute-Solution in the Presence of Detergents and Lipids. *Analytical Biochemistry* **1984,** *138* (1), 141-143.

62. Chen, W. X.; Smeekens, J. M.; Wu, R. H., A Universal Chemical Enrichment Method for Mapping the Yeast N-glycoproteome by Mass Spectrometry (MS). *Mol Cell Proteomics* **2014,** *13* (6), 1563-1572.

63. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

64. Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Apweiler, R.; Alpi, E.; Antunes, R.; Ar-Ganiska, J.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Chavali, G.; Cibrian-Uhalte, E.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Gane, P.; Cas-Tro, L. G.; Garmiri, P.; Hatton-Ellis, E.; Hieta, R.; Huntley, R.; Legge, D.; Liu, W. D.; Luo, J.; MacDougall, A.; Mutowo, P.; Nightin-Gale, A.; Orchard, S.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G. Y.; Rosanoff, S.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Turner, E.; Volynkin, V.; Wardell, T.; Watkins, X.; Watkins; Cowley, A.; Figueira, L.; Li, W. Z.; McWilliam, H.; Lopez, R.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; De Castro, E.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Nouspikel, N.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C. M.; Chen, Y. X.; Garavelli, J. S.; Huang, H. Z.; Laiho, K. T.; McGarvey, P.; Natale, D. A.; Suzek, B. E.; Vinayaka, C. R.; Wang, Q. H.; Wang, Y. Q.; Yeh, L. S.; Yerramalla, M. S.; Zhang, J.; Consortium, U., UniProt: a hub for protein information. *Nucleic Acids Res* **2015,** *43* (D1), D204-D212.

65. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009,** *37* (1), 1-13.

66. Lenz, L. L.; Mohammadi, S.; Geissler, A.; Portnoy, D. A., SecA2-dependent secretion of autolytic enzymes promotes Listeria monocytogenes pathogenesis. *P Natl Acad Sci USA* **2003,** *100* (21), 12432-12437.

67. Nuoffer, C.; Jeno, P.; Conzelmann, A.; Riezman, H., Determinants for glycophospholipid anchoring of the saccharomyces-cerevisiae GAS1 protein to the plasma-membrane. *Mol. Cell. Biol.* **1991,** *11* (1), 27-37.

68. Ghaemmaghami, S.; Huh, W.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S., Global analysis of protein expression in yeast. *Nature* **2003,** *425* (6959), 737-741.

CHAPTER 5. ENHANCING THE MASS SPECTROMETRIC IDENTIFICATION OF MEMBRANE PROTEINS

Adapted with permission from the Royal Society of Chemistry

Smeekens, J.M., Chen, W., Wu, R. Enhancing the mass spectrometric identification of membrane proteins by combining chemical and enzymatic digestion methods. *Analytical Methods*. 2015, 7 (17), 7220-7227.

5.1 Introduction

Membrane proteins play extremely important roles in biological systems and are crucial for a variety of cellular events including cell signaling, extracellular interactions and molecular transport.[1-3] They also participate in various cellular functions, including adhesion, growth and metastasis, which contribute to disease progression.[4-5] Additionally, because of their location on the cell surface, and resulting accessibility by macromolecules, membrane proteins are admirable for their potential as therapeutic and diagnostic targets.[6-7] It has been estimated that about one third of the genome encodes membrane proteins,[8] yet they represent 60-70% of FDA approved drug targets.[9-10] The comprehensive analysis of membrane proteins will facilitate a better understanding of membrane protein function and lead to the identification of membrane proteins as effective biomarkers and drug targets.[11] However, the hydrophobic nature and low abundance of membrane proteins hinders their global analysis.[12-14] Modern mass spectrometry (MS)-based proteomics techniques have proven to be very powerful for global protein analysis.[15-20] Common bottom-up proteomics techniques, where proteins are digested into peptides and subsequently analyzed with mass spectrometry,[21-23] require

effective digestion wherein proteins remain solubilized so that proteases or small molecules can access cleavage sites.[24] However, the accessibility of hydrophobic membrane proteins has been an existing problem because they tend to aggregate, precipitate and remain tightly folded in aqueous environments.[13]

Enzymatic methods are most commonly used for protein digestion prior to MS-based proteomics analysis. Several enzymes are frequently used, the most common being trypsin, which cleaves the peptide bond at the C-terminus of lysine and arginine residues (unless followed by a proline).[25] Trypsin is preferred due to the relatively high abundances of lysine and arginine, and their distribution throughout proteins,[26] which results in many peptides ideal for MS analysis. Other enzymes investigated here and frequently used for protein digestion include Lys-C and Glu-C. Lys-C cleaves at the C-terminus of lysine residues,[27] and Glu-C cleaves at the C-terminus of glutamic and aspartic acid residues.[28] Membrane protein digestion with only enzymatic methods is challenging since bulky enzymes cannot access all cleavage sites within tightly folded hydrophobic proteins. As a result, digestion exclusively with enzymes is often not sufficient to achieve comprehensive analysis of membrane proteins.

The inaccessibility of membrane proteins by proteases may be at least partially improved with chemical digestion methods, where small molecules can more easily access cleavage sites. However, chemical cleavage generally targets amino acids that are less abundant within proteins, which results in larger peptide fragments.[29] For example, cyanogen bromide (CNBr) targets methionine residues and 2-nitro-5-thiocyanobenzoic acid (NTCB) targets cysteine residues.[30] NTCB only cleaves reduced cysteine residues, so protein reduction is required before digestion can be performed. Digestion with NTCB

leads to two types of products, one is a N-terminal peptide and a cyclized N-terminal cysteine, and the other is dehydroalanine, which is the product of β-elimination on the thiocyanato group of cysteine.[31]

Theoretically, the combination of chemical and enzymatic methods would be ideal for membrane protein digestion because small molecules can easily access cleavage sites within membrane proteins, and the resulting fragments would be easily accessible by proteases. In this work, chemical and enzymatic methods were combined to improve membrane protein digestion for MS analysis; chemical digestion was first performed with NTCB, and followed by enzymatic digestion with Glu-C or Lys-C and trypsin. These combinatorial digestion methods were compared to sequential enzymatic digestion with Lys-C and trypsin. Our experimental results confirmed that the combinatorial digestion method utilizing NTCB, Lys-C and trypsin was most efficient for membrane protein digestion.

5.2 Experimental Methods

5.2.1 Chemicals and reagents

NTCB was purchased from Tokyo Chemical Industry, Co., trypsin was from Promega, Lysyl endopeptidase (Lys-C) was from Wako, and Endoproteinase Glu-C from *Staphylococcus aureus* was from EMD Millipore. All other reagents were purchased from Sigma Aldrich if not stated. Zirconia/silica beads (0.5 mm diameter) were purchased from BioSpec Products. HEK 293T cells were kindly provided by Dr. Gang Bao's research group at the Georgia Institute of Technology.

5.2.2 Cell culture, lysis and membrane protein enrichment

HEK 293T cells were grown in Dulbecco's Modified Eagle's Medium (DMEM) and harvested when they reached ~80% confluency. Cell pellets were washed three times with PBS. To remove cytosolic proteins, digitonin buffer (150 mM NaCl, 50 mM HEPES (pH=7.4), 25 µg/mL digitonin, protease inhibitor (1 tablet/10 mL)) was added to the cell pellet and incubated with end-over-end rotation at 4 °C for ten minutes. The suspended cell pellet was centrifuged at 2,000 g for 10 minutes. Digitonin buffer was added to the cell pellet and the sample was vortexed and subsequently centrifuged at 2,000 g for 10 minutes. This digitonin wash was repeated for a total of two washes.

Lysis buffer (10 mM HEPES (pH=8), 1.5 mM $MgCl_2$, 10 mM KCl) and zirconia/silica beads were added to the cell pellet, and placed in the Mini-Beadbeater (BioSpec). Samples were subjected to three 30 second cycles with 2 minutes of resting on ice in between. A flame-heated needle was used to poke holes in the bottom of the tubes, and the contents were transferred to new tubes (leaving the beads behind) through centrifugation at 1,000 g for 3 minutes. Samples were vortexed and centrifuged at 2,500 g for 5 minutes to remove cell debris. The supernatants were transferred to new tubes and centrifuged at 25,830 g for 30 minutes. Sodium carbonate buffer (0.1M sodium carbonate, 1 mM EDTA) was added to the cell pellets, vortexed, and incubated for 30 minutes on ice. The samples were centrifuged at 25,830 g for 15 minutes, and the supernatant was removed. Urea buffer (75 mM NaCl, 50 mM HEPES (pH=7.4), 8 M urea) was added to the cell pellets and incubated with shaking for 30 minutes at room temperature. The samples were centrifuged at 25,830 g for 15 minutes, the supernatants were removed, and the urea incubation was repeated once. After the samples were

centrifuged and supernatants were removed, 1% SDC in PBS was added to samples to solubilize membrane proteins. Samples were incubated overnight with end-over-end rotation at room temperature. The following day, samples were centrifuged at 15,000 g for 15 minutes; the supernatants were transferred to new tubes and the pellets were discarded. Disulfide bonds within proteins were subjected to reduction by incubation with 5 mM DTT for 25 minutes at 56 °C. The sample was then divided into three equal samples and transferred to 10 kDa filter columns.

5.2.3 Lys-C and trypsin digestion

One filter column was centrifuged at 10,000 g for 5 minutes. The alkylation reaction buffer (14 mM iodoacetamide, 0.1% SDS, PBS (pH=8)) was added, and the sample was centrifuged at 10,000 g for 5 minutes. Reaction buffer was added again and incubated in the dark for 20 minutes at room temperature. After incubation, the sample was centrifuged at 10,000 g for 5 minutes, washed with 100 mM sodium acetate (pH=5), centrifuged at 10,000 g for 5 minutes and rinsed with water twice. Digestion buffer (50 mM NH_4OAc, 5% ACN, 0.1 M urea) and Lys-C (enzyme:substrate ratio of ~1:100) were added to the sample and incubated overnight at 37 °C. The next day, trypsin was added at an enzyme:substrate ratio of ~1:100 and the sample was subsequently incubated for four hours and centrifuged at 10,000 g for 30 minutes. The flow-through was collected, and the filter column was washed with 50 mM ammonium bicarbonate (pH=8.5). The second flow-through was collected and combined with the first. Combined flow-throughs were acidified with TFA to a final pH of ~2, and centrifuged at 2,500 g for 5 minutes. The supernatant was desalted on a 50 mg tC18 SepPak cartridge and dried. Eluted peptides

were dissolved in 30 µL of MS solvent (5% ACN, 4% formic acid (FA)), and 4 µL were analyzed by LC-MS/MS.

5.2.4 NTCB and enzymatic digestion

The other two samples were used for combined chemical and enzymatic digestion. First, samples were centrifuged at 10,000 g for 5 minutes. Reaction buffer (10 mM NTCB, 0.1% SDS, PBS (pH=8)) was added to the filter columns which were subsequently centrifuged at 10,000 g for 5 minutes. The buffer was added again to columns and incubated in the dark for 20 minutes at 40 °C, and then centrifuged at 10,000 g for 5 minutes. Sodium acetate (100 mM, pH=5) was added to filter columns, which were then centrifuged at 10,000 g for 5 minutes. Filter columns were rinsed with water and centrifuged at 10,000 g for 5 minutes twice. Next, 50 mM ammonium acetate (pH=9) was added to filter columns and incubated for one hour at 50 °C. Filter columns were centrifuged at 10,000 g for 5 minutes and rinsed with water twice. To one filter column, 200 µL PBS (pH=7.4) and Glu-C (enzyme:substrate ratio of ~1:100) were added. To the second filter column, 200 µL digestion buffer (NH_4OAc (pH=8.5), 5% ACN, 0.1M urea) and Lys-C at an enzyme:substrate ratio of ~1:100 were added. Both samples were incubated overnight at 37 °C. The next day, trypsin was added (enzyme:substrate ratio of ~1:100) to the Lys-C digestion sample, which was subsequently incubated for four hours. Both digestion samples were centrifuged at 10,000 g for 30 minutes. The flow-through was collected for each, and filter columns were washed with 50 mM ammonium bicarbonate (pH=8.5). The second flow-through for each was collected and combined with the first. Combined flow-throughs were quenched with

TFA to a final pH of ~2. Acidified samples were centrifuged at 2,500 g for 5 minutes and the supernatant was desalted on a 50 mg tC18 SepPak cartridge and dried. Eluted peptides were dissolved in 30 µL of MS solvent, and 4 µL were analyzed by LC-MS/MS.

5.2.5 LC-MS/MS analysis

Samples were loaded onto the same HPLC as described in previous chapters, and peptides were separated with a 90 minute gradient of 3-25% ACN containing 0.125% FA. Samples were detected in the same LTQ Orbitrap Elite as described previously.

5.2.6 Database searches

The raw MS files were converted into mzXML format and searched using the SEQUEST algorithm (version 28)[32] against a database that included sequences of all proteins in the UniProt Human (*Homo sapiens*) Database and common contaminants. FDR were estimated at the peptide level as described in previous chapters. A 20 ppm precursor mass tolerance and 1.0 Da product ion mass tolerance were used in the database search and no enzyme was specified. Samples digested with NTCB were searched with a parameter file listing the following differential modifications: oxidation of methionine (+15.9949), β-elimination of cysteine (-33.9877), cyclized N-terminal cysteine (+24.9952); and one fixed modification: carbamidomethylation of cysteine (+57.0214). The sample digested exclusively with Lys-C and trypsin was searched using a parameter file that listed the oxidation of methionine (+15.9949) as a differential modification and the carbamidomehylation of cysteine (+57.0214) as a fixed modification.

5.2.7 Data filtering

Linear discriminant analysis (LDA) was utilized to distinguish correct and incorrect peptide identifications as described in previous chapters. After scoring, peptides less than six amino acids in length were discarded and peptide spectral matches were filtered to a less than 1% FDR at the peptide level based on the number of decoy sequences in the final data set.

5.3 Results and Discussion

5.3.1 Peptide and protein identification

Membrane proteins were extracted from HEK 293T cells as described above, and the membrane protein-enriched sample was equally split into three aliquots for subsequent digestion with different methods. The first sample was subjected to enzymatic digestion with Lys-C overnight and then trypsin for 4 hours (termed “LT” throughout this dissertation). Chemical and enzymatic methods were combined to perform the other digestions: one sample was digested with NTCB for 20 minutes and then Glu-C overnight (termed “NG”), and the other was digested with NTCB for 20 minutes, Lys-C overnight and then trypsin for 4 hours (termed “NLT”). These digested samples were purified with the stage-tip protocol and analyzed by LC-MS/MS (Figure 5.1).

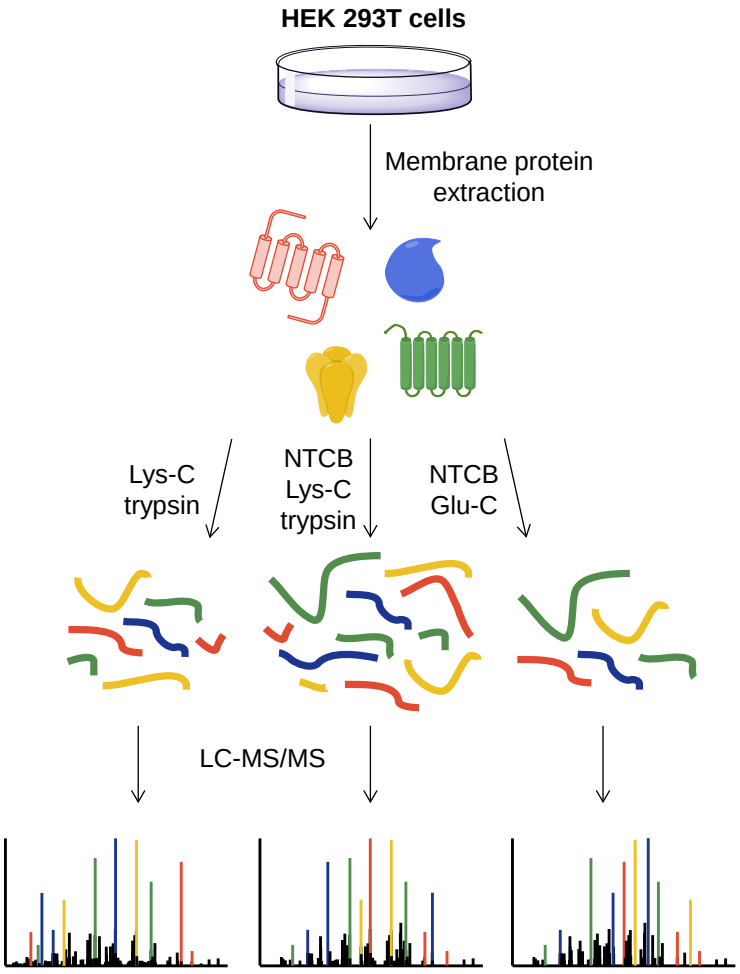

Figure 5.1. Experimental procedure comparing three digestion methods for the comprehensive analysis of membrane proteins.

Examples of mass spectra acquired from each digestion sample are shown in Figure 5.2. All three peptides are from HSPD1, a mitochondrial heat shock protein, which is involved in mitochondrial protein import and may play a subsequent role in the correct folding of imported proteins. This protein can be located in several subcellular components, including the extracellular region, cell surface and membrane.[33] The first peptide, DVANNTNEEAGDGTTTATVLAR, was identified from the sample digested

with the LT method with an XCorr of 5.1 and mass accuracy of -1.02 ppm. The second peptide, NAGVEGSLIVEK, identified with an XCorr of 3.5 and mass accuracy of -0.12 ppm, was from the NLT method. The third peptide, PLVIIAEDVDGEALSTLVLNRLK, was digested with the NG method and identified with an XCorr of 3.8 and mass accuracy of -0.67 ppm. All three of these peptides have XCorr values greater than 3, which indicates strong correlation between the mass spectra acquired in these experiments and the corresponding theoretical mass spectra. Furthermore, the mass accuracy associated with each of these peptides was also very high (within ± 2 ppm).

The combination of chemical and enzymatic digestion in the NLT method allowed the identification of 9,843 total peptides, corresponding to 2,120 proteins in the membrane protein-enriched sample. Among all proteins identified with the NLT method, 1,078 or 51% were membrane proteins. Proteins were identified as membrane proteins through cellular component clustering analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID).[34-35] Just over half of the proteins identified were membrane proteins, which is due to the fact that complete separation of membrane and non-membrane proteins remains challenging.[36] The work presented here focused mainly on the comparison of digestion methods, not complete coverage of the membrane proteome.

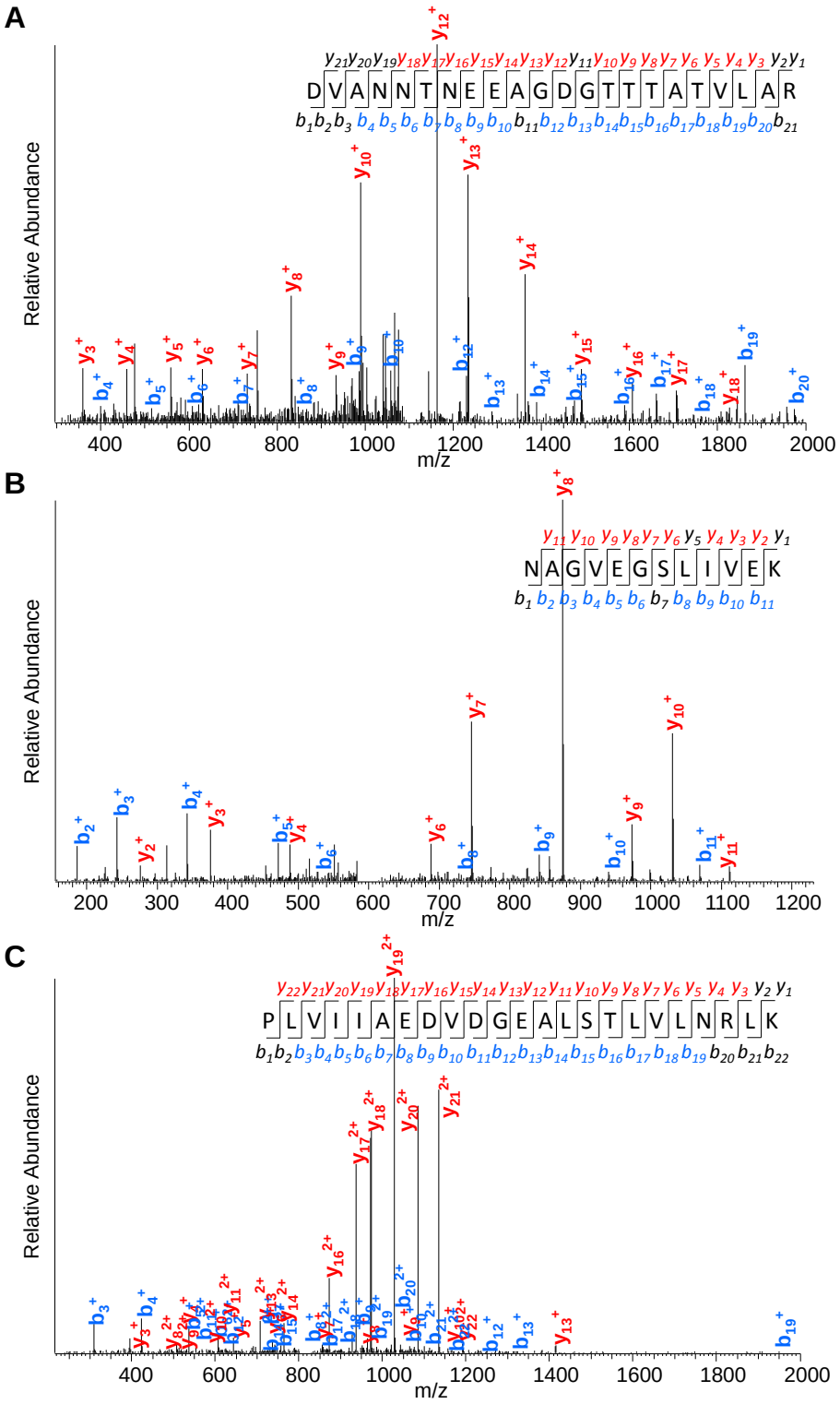

Figure 5.2. Tandem mass spectra corresponding to peptides identified from the protein HSPD1 using digestion methods combining (A) Lys-C and trypsin, (B) NTCB, Lys-C and trypsin, and (C) NTCB and Glu-C.

5.3.2 Comparison of three digestion methods

Compared to the LT method, which only utilized enzymatic digestion to identify 7,982 peptides, 23% more peptides were identified with the NLT method, corresponding to 20% more proteins (1,764 proteins found with LT). In an aqueous digestion environment, many cleavage sites targeted by Lys-C and/or trypsin may not be accessible by these two proteases. However, an initial chemical digestion by NTCB cleaved proteins into several fragments, allowing Lys-C and/or trypsin to access the appropriate cleavage sites. Additionally, small molecules like NTCB can more easily access cleavage sites within folded membrane proteins.

The other combinatorial digestion method employing NTCB and Glu-C (NG) provided the fewest number of total peptides (3,673) and proteins (1,037). Glutamic acid and aspartic acid are less abundant in proteins and therefore, fewer Glu-C cleavage sites exist; as a result, Glu-C is not as effective as Lys-C and/or trypsin. Additionally, the digestion efficiency of Glu-C is not as high as Lys-C and trypsin. These results are consistent with the fact that Lys-C and trypsin are much more frequently used in bottom-up proteomics.

In addition to total peptides and proteins identified, the number of membrane proteins identified in each experiment was also compared. The greatest number of membrane proteins, 1,078 proteins, was identified from the experiment that used the NLT method. Figure 5.3 shows the number of membrane proteins identified with each digestion method; a similar trend is seen among membrane proteins as total peptides and

proteins. All peptides identified with each digestion method are listed in a supplementary table online at doi.org/10.1039/C5AY00494B.

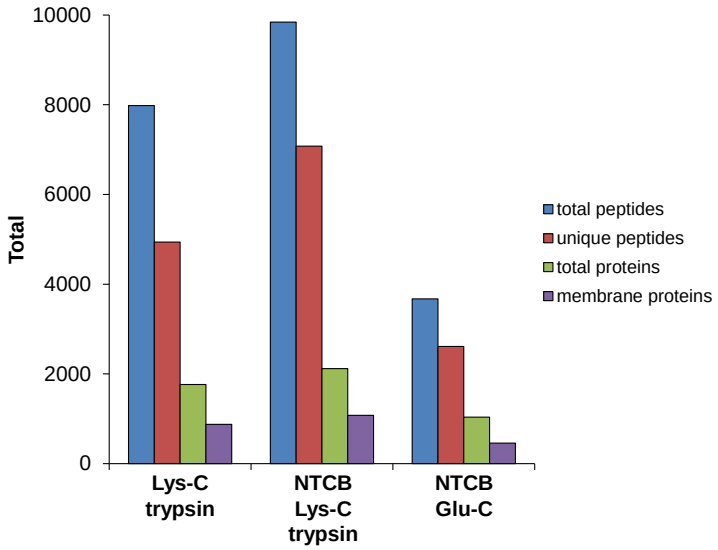

Figure 5.3. Number of total peptides, unique peptides, proteins and membrane proteins identified using each digestion method.

5.3.3 Peptide and protein overlap among three digestion methods

The overlap between peptides and proteins identified with each method was also investigated. Figure 5.4A shows the overlap between unique peptides identified with each digestion method; 109 peptides were identified in all samples. There was very little overlap between peptides identified in samples digested with NG and either of the other methods (170 peptides between NG and LT, and 253 between NG and NLT), which is expected due to the different cleavage site specificities of each enzyme, particularly between NG and LT which have no overlapping cleavage sites. In contrast, there is

significant peptide overlap between the LT and NLT experiments; a total of 2,391 unique peptides were identified with both methods. The protein overlap between all three samples is shown in Figure 5.4B, and the number of proteins identified with multiple methods is markedly higher. Peptides from the same proteins are expected to be identified in multiple experiments, even if the peptides differ in sequence. More than 70% of proteins identified in the NG experiment (773 of 1,037 proteins) were also identified in the NLT experiment, and 76% proteins identified with the LT method (1,341 of 1,764 proteins) were also found with the NLT method. The overlap between membrane proteins identified in each method is shown in Figure 5.4C; 700 membrane proteins were identified in both LT and NLT experiments, and 329 proteins were found in all three experiments. A total of 1,308 membrane proteins were identified in this work.

5.3.4 Missed cleavage and peptide length distributions

Datasets were further analyzed to determine the number of missed cleavages associated with each digestion method (Figure 5.5A). Both methods utilizing Lys-C and trypsin resulted in the highest percentages of zero missed cleavages: 94% for LT and 92% for NLT. As shown in Figure 5.5A, NG resulted in the greatest percentage of peptides containing missed cleavages (80% compared to 6% and 8% for LT and NLT, respectively), which further demonstrate that NG was not a robust digestion method. Although the method combining NTCB, Lys-C and trypsin gave a slightly lower percentage of zero missed cleavage sites compared to the LT method, 23% more total peptides and 20% more proteins were identified, as discussed above. Overall, the method combining NTCB, Lys-C and trypsin provided the most effective digestion of membrane proteins for MS analysis.

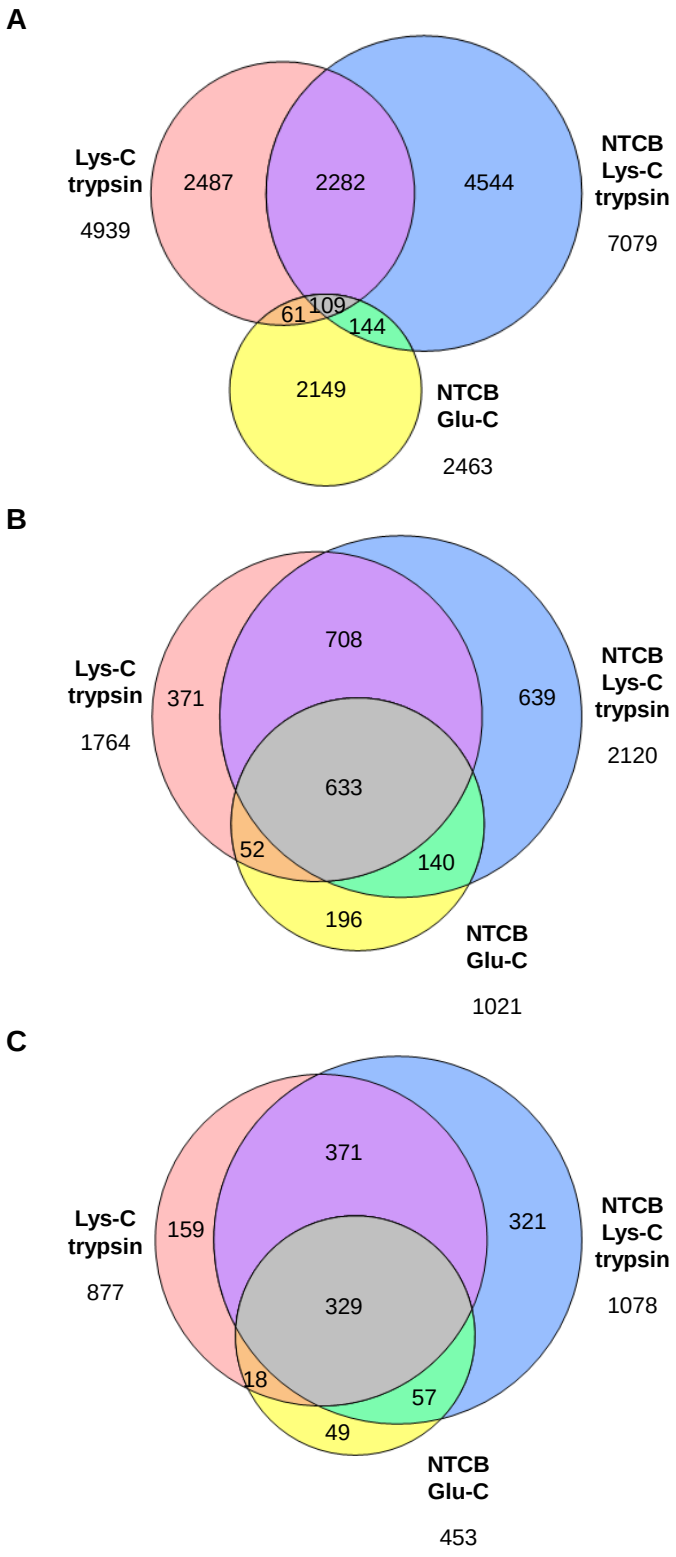

Figure 5.4. Overlap between (A) peptides, (B) proteins, and (C) membrane proteins identified using each of the three digestion methods.

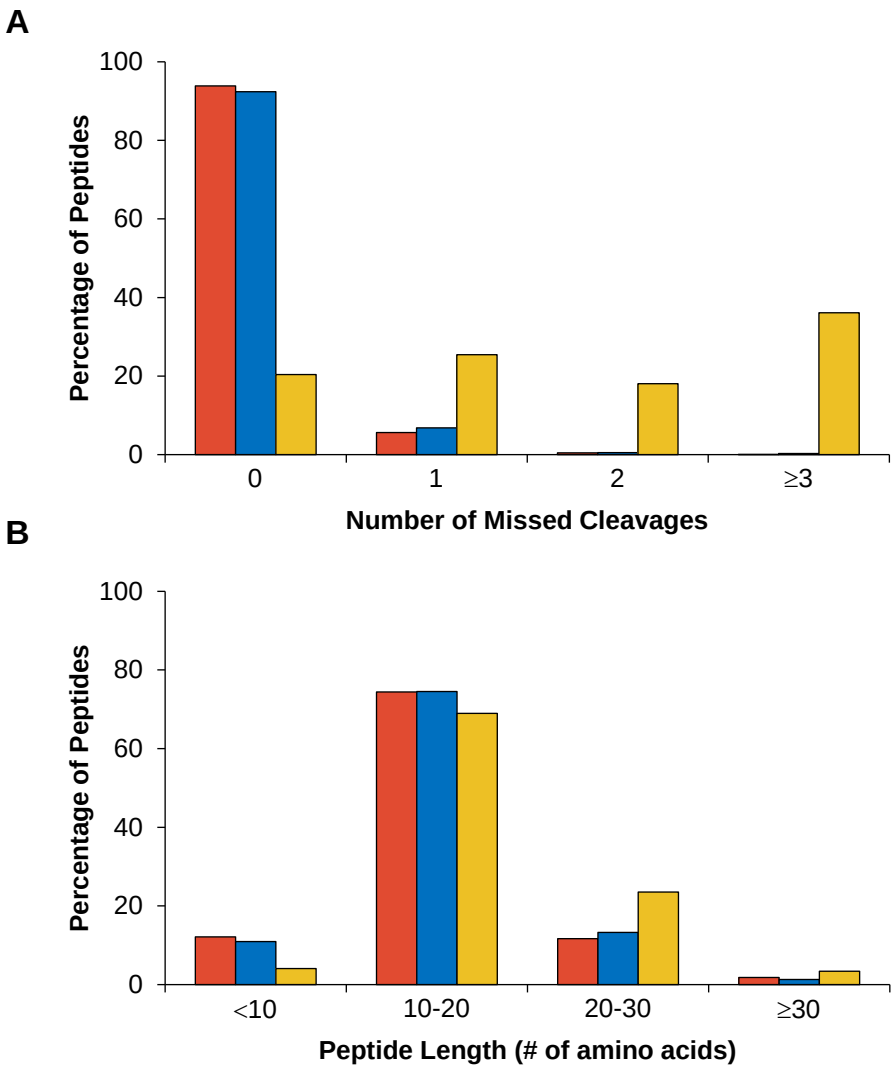

Figure 5.5. (A) Number of missed cleavages among peptides identified with each digestion method; (B) Distribution of peptide length for each digestion method. (Red: Lys-C and trypsin; Blue: NTCB, Lys-C and trypsin; Yellow: NTCB and Glu-C).

The length of peptides generated from each digestion method was also investigated. Figure 5.5B shows the peptide length distribution for each digestion method. There is no significant difference between methods, except that the overall distribution of peptide length in the NG experiment includes a greater number of larger peptides. Because chemical digestion was always used in conjunction with enzymatic

methods, the differences in peptide length typically seen between enzymatic and chemical cleavage could be compensated for with the sequential enzymatic methods. Additionally, MS is biased towards a specific peptide length range (10-30 amino acids), and based on the parameters used in these experiments, peptides with too few or too many amino acid residues may not be effectively detected.

5.3.5 Membrane protein clustering

Membrane proteins identified in samples digested with the NLT method were further studied through clustering analysis. Proteins were clustered according to biological process and molecular function using DAVID (Figure 5.6). Biological process clustering revealed that establishment of localization was most highly enriched with a *P* value of 4.5E-127. Proteins related to membrane organization and oxidation reduction were also highly enriched, and it is well-known that many oxidation and reduction reactions occur among membrane proteins in the mitochondria. A number of proteins with functions corresponding to cell adhesion and cell motion were also enriched. Molecular functions such as substrate-specific transporter activity, oxidoreductase activity, and protein binding were highly enriched among membrane proteins identified, which is consistent with known functions of membrane proteins.

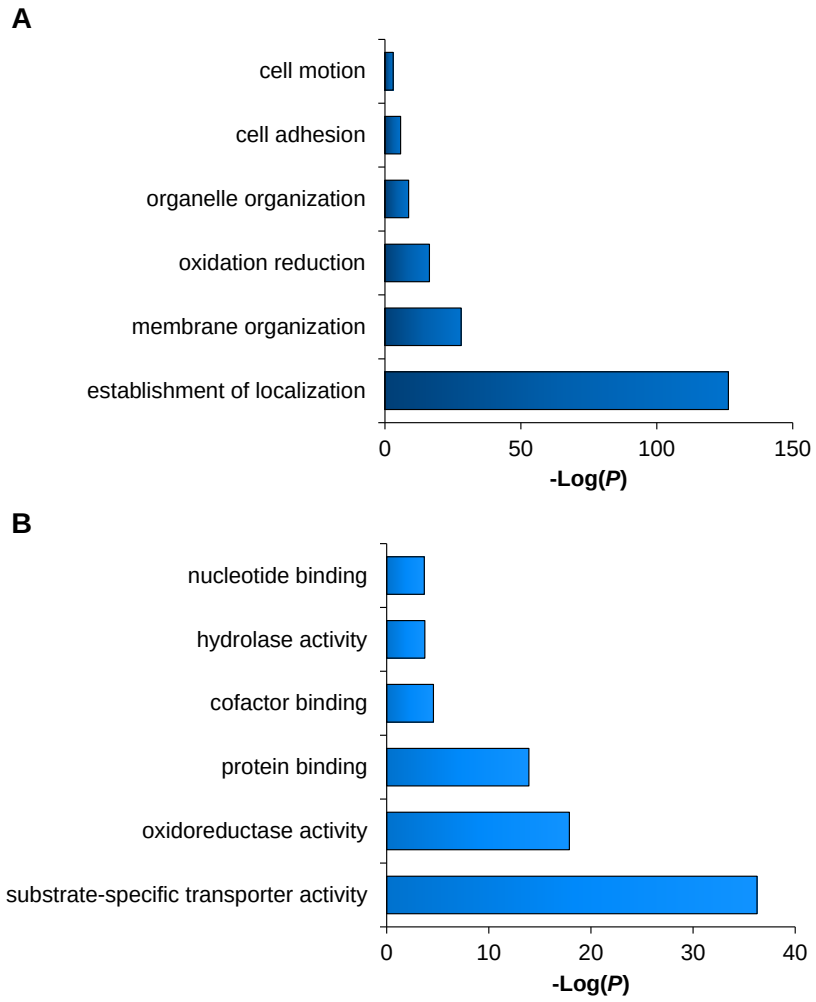

Figure 5.6. Clustering of membrane proteins identified in the NTCB, Lys-C and trypsin digestion sample according to (A) biological process and (B) molecular function.

5.3.6 SNARE complex proteins

One specific group of membrane proteins was further investigated here. The soluble *N*-ethylmaleimide-sensitive fusion protein attachment protein receptor (SNARE) complex assists in the fusion of transport vesicles with their targeted membranes.[37] They are often found on various membranes throughout the cell, and are essential for

intracellular membrane trafficking. Nine proteins in this complex were identified in this experiment: SNAP23, STX2, STX4, STX6, STX7, STX8, STX10, STX12, and VAMP2. For each protein, multiple unique peptides were identified, and several examples are listed in Table 5.1. The greatest number of total peptides was identified with the NLT method (77 peptides), and notable overlap exists between the peptides identified with the LT and NLT methods. Fewer total peptides were identified with the NG method (22 peptides), but they complement those identified with either the LT or NLT method. For example, a total of seven unique peptides were identified in the protein STX2, among which two peptides were identified with both the NLT and LT methods, four were identified with only the NLT method and one was identified with only the NG method. The combination of chemical and enzymatic methods demonstrated to be effective to analyze SNARE complex proteins, which play critical roles in membrane trafficking. All peptides identified from SNARE complex proteins are listed in a supplementary table online at doi.org/10.1039/C5AY00494B.

Table 5.1. Examples of SNARE proteins and peptides identified in this work.

Gene Symbol	Peptide	LT	NLT	NG
SNAP23	LNIGNEIDAQNPQIK	✓	✓	
	IQQRAHQITDE			✓
	RAHQITDESLESTR		✓	
	ILGLAIESQDAGIK	✓		
	ITNDAREDEMEENLTQVGSILGNLK		✓	
STX2	RSKGRIQRQLE			✓
	AIEQSFDQDESGNR		✓	
	NDDGDTVVVVEK	✓	✓	
	PSIFTSDIISDSQITR	✓	✓	
	FVETQGEMINNIER		✓	
STX4	KNILSSADYVE			✓
	VFVSNILKD			✓
	EEADENYNSVNTR	✓	✓	
	NILSSADYVER	✓		
	TQHGVLSQQFVELINK		✓	
STX6	MKDQMSTSSVQALAE			✓
	LLQDPSTATREE			✓
	AVNTAQGLFQR		✓	
	QALLGDSGSQNWSTGTTDK	✓	✓	
STX7	NVENAEVHVQQANQQLSR	✓	✓	
	FTTSLTNFQKVQRQAAE			✓
	QLGTPQDSPELR	✓	✓	
	NLVSWESQTQPQVQVQDEEITEDDLR		✓	
STX8	ALSSIISRQKQMGQE			✓
	IIQEQDAGLDALSSIISR		✓	
	GLGFDEIR	✓	✓	
	QNLLDDLVTR		✓	
STX10	ANPGKFKLPAGDLQE			✓
	ILAGKPAAQKSPSDLLDASAVSATSRYIEE			✓
	SPSDLLDASAVSATSR	✓	✓	
	VSGSIQVLK	✓		
STX12	QSQEDEVAITEQDLELIK		✓	
	ISQATAQIK	✓	✓	
	LMNDFSAALNNFQAVQR		✓	
	NPGPSGPQLR	✓		
	QLEADILDVNQIFK		✓	
VAMP2	IMRVNVDKVLE			✓
	LSELDDRADALQAGASQFETSAAK		✓	
	ADALQAGASQFETSAAK	✓	✓	
	LQQTQAQVDEVVDIMR	✓	✓	
	AATAPPAAPAGEGGPPAPPPNLTSNR		✓	

5.4 Conclusions

Membrane proteins are extremely important in biological systems due to their involvement in a variety of cellular processes, including signal transduction, molecular transport, cell-cell communication and cell-environment interactions. Membrane proteins are notoriously difficult to analyze, even with powerful modern MS-based proteomics techniques, because of their hydrophobicity and overall low abundance. This work presents combinatorial methods incorporating chemical and enzymatic digestion to cleave proteins for MS analysis. Parallel experiments clearly demonstrated that the combination of NTCB with Lys-C and trypsin can provide 23% more total peptides and 20% more proteins identified than the common Lys-C and trypsin digestion method. Chemical methods utilizing small molecules can more easily access cleavage sites within membrane proteins, compared to enzymatic methods which face steric hindrance. The combination of NTCB and Glu-C was not as effective, shown through the number of missed cleavages identified. Between the three digestion methods compared here, totally over 1,300 membrane proteins were identified. The combination of chemical and enzymatic methods demonstrated to be effective for membrane protein digestion, and further implementation of this method will allow the comprehensive and quantitative analysis of membrane proteins in complex biological samples.

5.5 References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., *Molecular Biology of the Cell*. Garland Science: New York, 2002.

2. Gilmore, J. M.; Washburn, M. P., Advances in shotgun proteomics and the analysis of membrane proteomes. *J. Proteomics* **2010,** *73* (11), 2078-2091.

3. Lee, J. K.; Stroud, R. M., Unlocking the eukaryotic membrane protein structural proteome. *Curr. Opin. Struct. Biol.* **2010,** *20* (4), 464-470.

4. Kampen, K. R., Membrane proteins: the key players of a cancer cell. *J Membrane Biol* **2011,** *242* (2), 69-74.

5. Macher, B. A.; Yen, T. Y., Proteins at membrane surfaces - a review of approaches. *Mol. Biosyst.* **2007,** *3* (10), 705-713.

6. Grimm, D.; Bauer, J.; Pietsch, J.; Infanger, M.; Eucker, J.; Eilles, C.; Schoenberger, J., Diagnostic and therapeutic use of membrane proteins in cancer cells. *Curr Med Chem* **2011,** *18* (2), 176-190.

7. Rucevic, M.; Hixson, D.; Josic, D., Mammalian plasma membrane proteins as potential biomarkers and drug targets. *Electrophoresis* **2011,** *32* (13), 1549-1564.

8. Wallin, E.; von Heijne, G., Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **1998,** *7* (4), 1029-1038.

9. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat Rev Drug Discov* **2002,** *1* (9), 727-730.

10. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M., Drug-target network. *Nat Biotechnol* **2007,** *25* (10), 1119-1126.

11. Weinglass, A. B.; Whitelegge, J. P.; Kaback, H. R., Integrating mass spectrometry into membrane protein drug discovery. *Curr Opin Drug Disc* **2004,** *7* (5), 589-599.

12. Eichacker, L. A.; Granvogl, B.; Mirus, O.; Muller, B. C.; Miess, C.; Schleiff, E., Hiding behind hydrophobicity - Transmembrane segments in mass spectrometry. *J Biol Chem* **2004,** *279* (49), 50915-50922.

13. Helbig, A. O.; Heck, A. J. R.; Slijper, M., Exploring the membrane proteome-Challenges and analytical strategies. *J. Proteomics* **2010,** *73* (5), 868-878.

14. Wei, X.; Dulberger, C.; Li, L. J., Characterization of murine brain membrane glycoproteins by detergent assisted lectin affinity chromatography. *Anal. Chem.* **2010,** *82* (15), 6329-6333.

15. Wu, C. C.; Yates, J. R., The application of mass spectrometry to membrane proteomics. *Nat Biotechnol* **2003,** *21* (3), 262-267.

16. Savas, J. N.; Stein, B. D.; Wu, C. C.; Yates, J. R., Mass spectrometry accelerates membrane protein analysis. *Trends Biochem.Sci.* **2011,** *36* (7), 388-396.

17. Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., A method for the comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol* **2003,** *21* (5), 532-538.

18. Zhao, Y. X.; Zhang, W.; Kho, Y. J.; Zhao, Y. M., Proteomic analysis of integral plasma membrane proteins. *Anal. Chem.* **2004,** *76* (7), 1817-1823.

19. Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M., Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **2010,** *7* (9), 681-685.

20. Wu, R. H.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Methods* **2011,** *8* (8), 677-683.

21. McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **1997,** *69* (4), 767-776.

22. Aebersold, R.; Goodlett, D. R., Mass spectrometry in proteomics. *Chem. Rev.* **2001,** *101* (2), 269-295.

23. Sun, L. L.; Hebert, A. S.; Yan, X. J.; Zhao, Y. M.; Westphall, M. S.; Rush, M. J. P.; Zhu, G. J.; Champion, M. M.; Coon, J. J.; Dovichi, N. J., Over 10000 peptide identifications from the HeLa proteome by using single-shot capillary zone electrophoresis combined with tandem mass spectrometry. *Angew. Chem.-Int. Edit.* **2014,** *53* (50), 13931-13933.

24. Nagaraj, N.; Lu, A. P.; Mann, M.; Wisniewski, J. R., Detergent-based but gel-free method allows identification of several hundred membrane proteins in single LC-MS runs. *J Proteome Res* **2008,** *7* (11), 5028-5032.

25. Olsen, J. V.; Ong, S. E.; Mann, M., Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **2004,** *3* (6), 608-614.

26. Brownridge, P.; Beynon, R. J., The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods* **2011,** *54* (4), 351-360.

27. Raijmakers, R.; Neerincx, P.; Mohammed, S.; Heck, A. J. R., Cleavage specificities of the brother and sister proteases Lys-C and Lys-N. *Chem Commun* **2010,** *46* (46), 8827-8829.

28. Drapeau, G. R.; Houmard, J.; Boily, Y., Purification and properties of an extracellular protease of *Staphylococcus-aureus*. *J Biol Chem* **1972,** *247* (20), 6720-&.

29. Crimmins, D. L., Mische, S. M. and Denslow, N. D. , Chemical cleavage of proteins in solution. *Current Protocols in Protein Science* **2005,** (Chapter 11: Unit 11.4).

30. Tang, H. Y.; Speicher, D. W., Identification of alternative products and optimization of 2-nitro-5-thiocyanatobenzoic acid cyanylation and cleavage at cysteine residues. *Anal Biochem* **2004,** *334* (1), 48-61.

31. Jacobson, G. R.; Schaffer, M. H.; Stark, G. R.; Vanaman, T. C., Specific chemical cleavage in high-yield at amino peptide-bonds of cysteine and cystine residues. *J Biol Chem* **1973,** *248* (19), 6583-6591.

32. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

33. Consortium, T. U., Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **2014,** *42* (D1), D191-D198.

34. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4* (1), 44-57.

35. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009,** *37* (1), 1-13.

36. Lin, Y.; Wang, K. B.; Yan, Y. J.; Lin, H. Y.; Peng, B.; Liu, Z. H., Evaluation of the combinative application of SDS and sodium deoxycholate to the LC-MS-based shotgun analysis of membrane proteomes. *J. Sep. Sci.* **2013,** *36* (18), 3026-3034.

37. Brunger, A. T., Structure and function of SNARE and SNARE-interacting proteins. *Q. Rev. Biophys.* **2005,** *38* (1), 1-47.

CHAPTER 6. CONCLUSIONS AND FUTURE OUTLOOK

6.1 Mass spectrometric analysis of the cell surface N-glycoproteome

6.1.1 Summary of results

Cell surface N-glycoproteins play extraordinarily important roles in cell-cell communication, cell-matrix interactions and cellular response to environmental cues. Global analysis is exceptionally challenging because many N-glycoproteins are present at low abundances and effective separation is difficult to achieve. Here we have developed a novel strategy integrating metabolic labeling, copper-free click chemistry and MS-based proteomics methods to analyze cell surface N-glycoproteins comprehensively and site-specifically. A sugar analogue containing an azido group, N-azidoacetylgalactosamine, was fed to cells to label glycoproteins. Glycoproteins with the functional group on the cell surface were then bound to dibenzocyclooctyne-sulfo-biotin via copper-free click chemistry under physiological conditions. After protein extraction and digestion, glycopeptides with the biotin tag were enriched by NeutrAvidin conjugated beads. Enriched glycopeptides were deglycosylated with PNGase F in heavy-oxygen water, and in the process of glycan removal, asparagine was converted to aspartic acid and tagged with ^{18}O for MS analysis.

With this strategy, 144 unique N-glycopeptides containing 152 N-glycosylation sites were identified in 110 proteins in HEK 293T cells. As expected, 95% of identified glycoproteins were membrane proteins, which were highly enriched. Many sites were located on important receptors, transporters, and CD proteins. The experimental results

demonstrated that the current method is highly effective for the comprehensive and site-specific identification of the cell surface N-glycoproteome and can be extensively applied to other cell surface protein studies.

6.1.2 Future directions

The method presented here provides the foundation for a variety of cell surface glycoprotein studies. One of the main advantages of this method is that metabolic labeling and the click reaction to tag surface glycoproteins is performed under physiological conditions, allowing broad applications for investigating surface glycoprotein response to different stimuli. It can be applied extensively to comprehensively investigate surface glycoproteins in cultured cells or animal models, including mice and zebrafish. One general application is to investigate changes in surface glycoproteins as a result of drug treatment, including inhibitors or other stimuli. An example of this application is presented in Chapter 3, by pairing this method with quantitative proteomics to identify and quantify cell surface N-glycoproteins throughout epithelial-mesenchymal transition. Global analysis of cell surface glycoproteins will provide a catalog of proteins from which biomarkers and drug targets can be identified. Glycosylation site occupancy can also be exploited to gain a better understanding of cellular mechanisms related to disease and development.

6.2 Quantification of cell surface N-glycoproteome throughout epithelial-mesenchymal transition

6.2.1 Summary of results

By coupling TMT labeling with our cell surface N-glycoprotein method described in Chapter 2, surface glycoprotein abundance changes were quantified in MCF 10A cells undergoing the epithelial-mesenchymal transition. MCF 10A cells were treated with TGF-β to induce EMT, and cells were sampled at varying times (0, 4 and 8 days) to determine surface glycoprotein abundance changes throughout the transition. A total of 438 unique glycopeptides corresponding to 235 glycoproteins were quantified: 39 were down-regulated and 37 were up-regulated throughout EMT. Abundance changes throughout EMT were also investigated at the protein level: 4656 proteins were quantified, including 534 down-regulated and 699 up-regulated. Interestingly, down-regulated proteins were very highly enriched with biological processes associated with ribonucleoprotein complex biogenesis (P = 2.98E-62) and molecular functions related to RNA binding (P = 7.69E-87), implying that translation may be modulated during EMT. Up-regulated proteins were highly enriched with transport and oxidation-reduction processes. These results provide insight into the correlation between surface glycosylation and the molecular mechanisms of EMT.

6.2.2 Future directions

While the current results show interesting abundance changes in surface glycoproteins throughout EMT, the conclusions are somewhat limited by the time points studied. Further experiments can investigate more time points throughout the transition, to gain a more thorough understanding of when and how surface glycoprotein expressions change. To confirm the current results, the reverse process, mesenchymal to epithelial transition, can be induced and glycoprotein abundances can be compared. Based on the results that the majority of surface glycoproteins are not consistently up- or

down-regulated throughout EMT, future work can study the glycan structures and how they vary during the same time points. Since the glycans on these glycoproteins are the main players involved in extracellular interactions, it is possible that their structures and abundances are more dynamic than the proteins themselves. Investigating surface glycoprotein changes throughout EMT will provide a better understanding of the mechanisms of this transition and insight into how glycosylation and protein expression is modulated during the transition. Determining surface proteins that change in specific cancer or diseased cells may lead to the discovery of drug targets that would inhibit mesenchymal cell formation.

6.3 Global analysis of secreted proteins and glycoproteins in *Saccharomyces cerevisiae*

6.3.1 Summary of results

Protein secretion is essential for numerous cellular activities, and secreted proteins in bodily fluids are a promising and non-invasive source of biomarkers for disease detection. Systematic analysis of secreted proteins and glycoproteins will provide insight into protein function and cellular activities. Yeast (*Saccharomyces cerevisiae*) is an excellent model system for eukaryotic cells, but global analysis of secreted proteins and glycoproteins in yeast is challenging due to the low abundances of secreted proteins and contamination from high-abundance intracellular proteins. Here, by using mild separation of secreted proteins from cells, we comprehensively identified secreted proteins and glycoproteins and quantified their abundance changes through inhibition of protein glycosylation and MS-based proteomics.

In biological triplicate experiments, 245 secreted proteins were identified, and comparison with previous experimental and computational results demonstrated that many identified proteins were located in the extracellular space. The majority of quantified secreted proteins were down-regulated from cells treated with an N-glycosylation inhibitor (tunicamycin). The quantitative results strongly suggest that the secretion of these down-regulated proteins was regulated by glycosylation while, the secretion of proteins with minimal abundance changes was contrarily irrelevant to protein glycosylation, likely being secreted through non-classical pathways. Glycoproteins in the yeast secretome were globally analyzed for the first time. A total of 27 proteins were quantified in at least two protein and glycosylation triplicate experiments, and all except one were down-regulated under N-glycosylation inhibition, which is solid experimental evidence to further demonstrate that the secretion of these proteins is regulated by their glycosylation. These results provide valuable insight into protein secretion, which will further advance protein secretion and disease studies.

6.3.2 Future directions

This work provides preliminary information regarding protein secretion and its regulation by glycosylation. There are many avenues to pursue an in-depth understanding of this relationship, including beginning to distinguish between classical and non-classical secretion. Currently, the mechanisms behind non-classical secretion are ambiguous, which makes the differentiation of classical and non-classical secretion difficult in this study. One approach includes using another drug or inhibitor to prevent non-classical secretion to validate the group of proteins identified here that are believed to be secreted via non-classical pathways. Furthermore, glycan structures and maturity

can be investigated to determine how the glycan modifications are impacted when secretion is blocked. Treating cells with castanospermine, a glucosidase inhibitor, will limit N-glycan trimming and subsequently prevent glycan structures from reaching their correct final structure. These results may reveal specific structures that induce or prevent glycoprotein trafficking and secretion. Understanding protein secretion on a comprehensive scale will provide critical information regarding cellular processes. Furthermore, secreted proteins in bodily fluids represent a promising source of non-invasive biomarkers for disease detection and surveillance.

6.4 Enhancing the mass spectrometric identification of membrane proteins

6.4.1 Summary of results

Membrane proteins are critical for many cellular events, including cell signaling, molecular transport, and extracellular interactions. One third of the genome is estimated to encode membrane proteins, which are correlated with disease progression and can serve as promising biomarkers and drug targets. Modern MS-based proteomics techniques facilitate the global analysis of proteins in complex biological samples; however, the hydrophobicity of membrane proteins inhibits their comprehensive analysis. Since membrane proteins are not easily accessible by proteases in aqueous solutions, a combinatorial method incorporating chemical and enzymatic digestion is presented here to improve the membrane protein digestion for MS analysis. Chemical digestion with 2-nitro-5-thiocyanatobenzoic acid (NTCB) was supplemented with enzymatic digestion (Glu-C, or Lys-C and trypsin) to determine the optimal combination of digestion methods.

Three parallel experiments were performed with membrane protein extracts from HEK 293T cells, and the results demonstrated that combining NTCB with Lys-C and trypsin resulted in the greatest number of total peptides (9,483 peptides). Comparatively, digestion with only Lys-C and trypsin allowed the identification of 7,982 total peptides, and sequential digestion with NTCB and Glu-C resulted in 3,673 peptides. By integrating chemical digestion before enzymatic digestion, NTCB could more easily access cleavage sites within membrane proteins, and the resulting peptide fragments were thus more accessible by proteases. The combination of chemical and enzymatic digestion presented here proved to be effective for membrane protein analysis.

6.4.2 Future directions

Combining chemical and enzymatic digestion increased the MS-based identification of membrane proteins compared to common enzyme-only digestion. However, the scope of this work could be expanded to include more proteases and chemicals to further optimize digestion. The overall workflow could be simplified through use of different chemical digestion, such as cyanogen bromide (CNBr), which does not require protein reduction prior to digestion. However, this may be at the cost of lower membrane protein identification, and should be investigated. This method can also be applied to membrane and surface glycoprotein studies, to enhance the proteome coverage. Improved membrane protein identification by mass spectrometry can be applied to discover potential drug targets for a variety of diseases.

APPENDIX A. COLLABORATIVE RESEARCH

This appendix briefly describes collaborative projects completed in conjunction with the work presented in this dissertation. Collaborations outside of the Wu group are described first, followed by collaborative projects within the Wu group.

A.1 Competitive protein binding influences heparin-based modulation of spatial growth factor delivery for bone regeneration

Hettiaratchi, M.H., Chou, C., Servies, N., Smeekens, J.M., Cheng, A., Essancy, C., Wu, R., McDevitt, T.C., Guldberg, R.E., Krishnan, L. Competitive protein binding influences heparin-based modulation of spatial growth factor delivery for bone regeneration, Tissue Engineering Part A, 2017, DOI:10.1089/ten.TEA.2016.0507. Copyright 2017 Mary Ann Liebert, Inc.

A.1.1 Summary of project

Tissue engineering strategies involving the in vivo delivery of recombinant growth factors are often limited by the inability of biomaterials to spatially control diffusion of the delivered protein within the site of interest. The poor spatiotemporal control provided by porous collagen sponges, which are used for the clinical delivery of bone morphogenetic protein-2 (BMP-2) for bone regeneration, have necessitated the use of supraphysiological protein doses, leading to inflammation and heterotopic ossification. This study describes a novel tissue engineering strategy to spatially control rapid BMP-2 diffusion from collagen sponges in vivo by creating a high affinity BMP-2 sink around the collagen sponge. We designed an electrospun polycaprolactone nanofiber mesh

containing physically entrapped heparin microparticles, which have been previously demonstrated to bind and retain large amounts of BMP-2. Nanofiber meshes containing 0.05 and 0.10 mg of microparticles/cm2 demonstrated increased BMP-2 binding and decreased BMP-2 release in vitro compared to meshes without microparticles. However, when microparticle-containing meshes were used in vivo to limit diffusion of BMP-2 delivered using collagen sponges in a rat femoral defect, no differences in heterotopic ossification or biomechanical properties were observed. Further investigation revealed that, although BMP-2 binding to heparin microparticles was rapid, the presence of serum components attenuated microparticle-BMP-2 binding and increased BMP-2 release in vitro. These observations provide a plausible explanation for the results observed in vivo and suggest that competitive protein binding in vivo may hinder the ability of affinity-based biomaterials to modulate growth factor delivery.

A.1.2 MS-based proteomics contributions

In order to identify serum proteins bound to microparticles, mass spectrometry analysis was performed on microparticles (1 mg) loaded with 5 mL of FBS or 5 mL of FBS and 8 μg BMP-2 together. The microparticles were centrifuged, washed once with PBS, and proteins bound to microparticles were digested with 2 units of glutamyl endopeptidase (Glu-C) for 16 hours and 10 units of lysyl endopeptidase (Lys-C) for 3.5 hours in 50 mM HEPES (pH=7.9). Eluted peptides were purified, dried, and resuspended in a 5% acetonitrile and 4% formic acid solution for LC-MS/MS analysis.

Mass spectrometry was used to identify serum proteins on FBS-loaded heparin microparticles. Proteins identified based on only two or fewer unique peptides were

excluded from the analysis. Out of 33 total proteins identified, 20 were known heparin-binding proteins, including several apolipoproteins and thrombospondins. Selected peptides identified from these heparin-binding proteins are highlighted in Table A.1. XCorr signifies the correlation between the experimental mass spectra and theoretical mass spectra of the peptide, while the mass accuracy of each peptide demonstrates how closely the theoretical mass of the peptide matches the experimentally measured mass.

Table A.1. Selected peptides from known heparin binding proteins identified on FBS-loaded heparin microparticles.

Protein	Peptide	ppm	XCorr
Apolipoprotein-A1	NWDTLASTLSK	1.03	3.3
	TLRQQLAPYSDD	0.63	2.9
	QLGPVTQE	0.27	2.6
Apolipoprotein-B	TSRSLPYAQNIQDQLSGLQE	1.30	5.9
	VSDSLIGVTQGYSVTVK	-0.28	4.5
	ITVPASQLTVSQFTLPK	0.77	4.1
Apolipoprotein-C2	TYLPAVDEK	0.55	2.4
	QVFSVLSGKD	0.71	2.0
	SLLGYWDTAK	0.83	2.9
Apolipoprotein-E	LQAAQARLGSDME	0.64	4.6
	YLRWVQTLSDQVQE	-0.06	4.3
	QGQSRAATLSTLAGQPLLE	0.73	3.8
Thrombospondin-1	DHSGQVFSVISNGK	-0.37	4.8
	FQDLVDAVRAE	0.45	3.6
	GPDPSSPAFRIE	0.83	3.1
Thrombospondin-4	FQTQNFDRLD	-0.13	2.4
	SSATIFGLYSSADHSK	0.70	2.3
	SSATIFGLYSSADHSK	0.04	3.8
Fibronectin-1	LGVRPSQGGEAPRE	0.62	4.4
	SLPLVGQQSTVSDVPRDLE	0.04	4.2
	YVYTISVLRDGQE	-0.21	4.2

The extracellular matrix protein fibronectin, which has a known heparin binding site and specific affinity for heparin, was confidently identified based on 26 unique peptides.[1] On microparticles loaded with both FBS and BMP-2, BMP-2 was only identified based on one unique peptide and thus was not included in the list of detected proteins. This further suggests that FBS borne proteins bound to the microparticles in larger quantities than BMP-2 and corroborates the results obtained from SDS-PAGE analysis.

A.2 Yeast rRNA Expansion Segments: Folding and Function

Gómez Ramos, L.M., Smeekens, J.M., Kovacs, N.A., Bowman, J.C., Wartell, R.M., Wu, R., Williams, L.D. Yeast rRNA expansion segments: Folding and function, *Journal of Molecular Biology*, 2016, 428 (20), 4048-4059. Copyright 2016 Elsevier.

A.2.1 Summary of project

Divergence between prokaryotic and eukaryotic ribosomal RNA (rRNA) and among eukaryotic ribosomal RNAs is focused in expansion segments (ESs). Eukaryotic ribosomes are significantly larger than prokaryotic ribosomes partly because of their ESs. We hypothesize that larger rRNAs of complex organisms could confer increased functionality to the ribosome. Here, we characterize the binding partners of *Saccharomyces cerevisiae* expansion segment 7 (ES7), which is the largest and most variable ES of the eukaryotic large ribosomal subunit and is located at the surface of the ribosome. In vitro RNA-protein pull-down experiments using ES7 as a bait indicate that ES7 is a binding hub for a variety of non-ribosomal proteins essential to ribosomal

function in eukaryotes. ES7-associated proteins observed here cluster into four groups based on biological process, (i) response to abiotic stimulus (e.g., response to external changes in temperature, pH, oxygen level, etc.), (ii) ribosomal large subunit biogenesis, (iii) protein transport and localization, and (iv) transcription elongation. Seven synthetases, Ala-, Arg-, Asp-, Asn-, Leu-, Lys- and TyrRS, appear to associate with ES7. Affinities of AspRS, TyrRS and LysRS for ES7 were confirmed by in vitro binding assays. The results suggest that ES7 in *S. cerevisiae* could play a role analogous to the multi-synthetase complex present in higher order organisms and could be important for the appropriate function of the ribosome. Thermal denaturation studies and footprinting experiments confirm that isolated ES7 is stable and maintains a near-native secondary and tertiary structure.

A.2.2 MS-based proteomics contributions

S. cerevisiae cell lysates were screened for proteins with affinity for ES7 using pull-down assays. Briefly, isolated ESs, or fragments of ESs, were attached to beads, incubated with cell lysates and washed to remove non-specific and weakly associated proteins. The remaining ES-associated proteins were proteolytically digested and the resulting peptides were identified by LC-MS/MS and analyzed by tandem mass spectra. During exponential growth of *S. cerevisiae*, 36 proteins were seen to interact with ES7. Our assay identifies primary proteins that interact directly with ES7 in addition to secondary proteins that bind to the primary proteins. We attached isolated ES7 to beads, incubated the beads with cell lysates, washed the beads to remove non-specific and weakly associated proteins, and then proteolytically digested the ES7-associated proteins.

Many of the proteins seen here to associate with ES7 have been previously shown to bind to the ribosome. Of the 36 ES7-associated proteins here, 25 were identified by Link as components of intact translation complexes.[2] These proteins include aspartyl tRNA synthetase (AspRS), translation machinery-associated protein 10 (TMA10), ribosome biogenesis protein ERB1 and ATP-dependent RNA helicase MAK5. Of these, 12 of them were identified as mRNA-binding proteins (mRBPs) by Hentze.[3] Nine proteins that associate with ES7 in this work were previously found to associate both with intact ribosomes and with mRNA.[2-3] Those nine are (i) protein transport factor SEC1; (ii) MAK21, which is involved in maturation and export of pre-LSU particles; (iii) fimbrin, which is an actin-bundling protein involved in cytoskeleton organization and maintenance; (iv) HXK1, which catalyzes hexose phosphorylation; (v) NAD(+) salvage pathway component, nicotinamidase; (vi) transcription elongation factor SPT6; (vii) AspRS; (viii) 40S ribosomal protein S30; and (ix) protein PBI2, inhibitor of vacuolar protein B.

Seven aminoacyl tRNA synthetases (aaRSs) were identified here in association with ES7. AspRS and tyrosyl tRNA synthetase (TyrRS) were found in each of the replicate experiments. Cumulatively Ala-, Arg-, Asn-, Asp-, Leu-, Lys- and TyrRS were observed in at least one experiment. Our observations are consistent with Link et al., who observed Ala-, Asn-, Asp-, Leu- and LysRS bound to intact ribosomes.[2] AspRS, LeuRS, and LysRS associate more tightly with the intact ribosome than AlaRS and AsnRS. The binding of AspRS, TyrRS and LysRS to ES7 is confirmed here by in vitro binding assays.

A total of six proteins known to localize in the nucleolus are observed here to associate with ES7.[4] Nucleolar proteins that associate with ES7 include (i) MAK5, (ii) MAK21, (iii) ERB1, (iv) ribosome biogenesis ATPase RIX7, (v) DNA-directed RNA polymerase I subunit RPA49 and (vi) RNA 3′-terminal phosphate cyclase-like protein RCL1.

Clustering of ES7-associated proteins by biological process was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID). Four clusters are highly enriched (Figure A.1 and Table A.2), including (i) response to abiotic stimulus such as temperature, (ii) ribosomal LSU biogenesis, (iii) protein transport and localization, and (iv) transcription elongation. For this analysis, if multiple clusters contained identical proteins, only the cluster with the lowest *P*-value was retained. The two most intense clusters correspond to response to abiotic stimulus ($P = 5.3 \times 10^{-3}$) and ribosomal LSU biogenesis ($P = 1.03 \times 10^{-2}$). With DAVID we were unable to cluster all 36 ES7-associated proteins because the functional annotation for the *S. cerevisiae* proteome is not fully finalized in the database. Of the 36 ES7-associated proteins, 12 were not clustered. Six mitochondrial proteins observed were assumed to be artifacts and were excluded from the analysis.

For comparison, we clustered Hentze's mRBPs by the same methods we used here for ES7-associated proteins, and we observed some of the same clusters. Clusters common to mRBPs and ES7-associated proteins include (i) response to abiotic stimulus and (ii) protein transport and localization. ES7-associated proteins were also clustered based on molecular function with DAVID (Figure A.1 and Table A.2). These results

demonstrate that ES7-associated proteins are important in (i) nucleotide binding and (ii) catalytic activity. Both of these clusters are also seen in mRBPs.

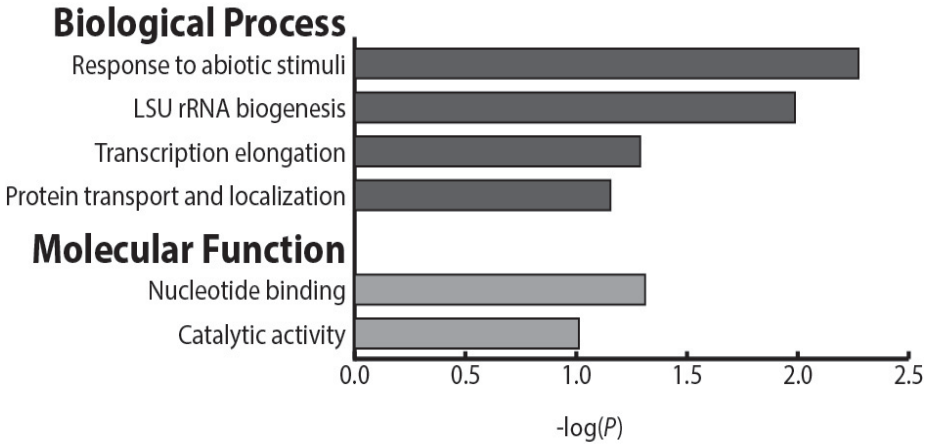

Figure A.1. Clustering of *S. cerevisiae* ES7-associated proteins identified in this work, based on biological process and molecular function.

Table A.2. List of genes included in each cluster in Figure A.1.

Biological Process	**Genes**
Response to abiotic stimuli	SEC1, PBS2, PBI2, SAC6, HXK1, TMA10, BDH1, PNC1
LSU rRNA biogenesis	MAK5, RIX7, ERB1, MAK21
Transcription elongation	SPT6, CHD1, CTR9
Protein transport and localization	NUM1, SEC1, PBS2, RIX7, YHB1, HXK1, YRB1, RTS1, YGR130C, VMA5, PEP8, SAC6, PBI2, SEC17
Molecular Function	**Genes**
Nucleotide binding	VMA5, MAK5, RIX7, PBS2, CHD1, HXK1, TYS1, DPS1, MET5, ERG1
Catalytic activity	MAL32, MAK5, PBS2, RIX7, SPT6, YHB1, HXK1, TYS1, PNC1, DPS1, ERG1, RTS1, RCL1, RPA49, VMA5, CHD1, GPT2, BDH1, MET5

A.3 Elongated expansion segments extend the capabilities of human ribosomes

Reproduced from

Gómez Ramos, L.M., Smeekens, J.M., Kovacs, N.A., Fang, P.Y., Petrov, A.S., Bernier, C.R., Hu, M., Bowman, J.C., Wartell, R.M., Wu, R., Williams, L.D. Elongated expansion segments extend the capabilities of human ribosomes. *Submitted.*

A.3.1 Summary of project

Metazoan ribosomes are elaborated by large expansion segments, which contain long unbranched helices marked by numerous bulges and mismatches. This work investigates the functions of the largest expansion segments (ESs) of the human ribosome using pull-down assays analyzed by mass spectroscopy. Human ESs associate with a variety of critical proteins including members of the ubiquitin-proteasome system. The results suggest that protein synthesis is physically coupled with protein degradation in humans. A variety of other functional proteins associate with ESs, including proteins involved in transcription, translation, signaling and structure. The distribution of ES-associated proteins is different for a non-carcinoma and a carcinoma cell line. We characterize the stability, secondary structure and in vitro folding of the largest human ES. Human ESs contain sequences with G-scores > 38 suggestive of potential G-quadruplexes. Some single-stranded ES fragments show K+-dependence of circular dichroism spectra, characteristic of G-quadruplex formation.

A.3.2 MS-based proteomics contributions

A.3.2.1 Protein binding to human rRNA expansion segments

The goal here is to characterize interactions between human ES rRNAs and human proteins. Human cell lysates were screened to identify proteins and protein complexes that associate with ES7 and ES27. To determine if protein-rRNA interactions differ based on cell line, multiple cell lines were investigated. The two cell lines investigated here are a non-carcinoma cell line (HEK 293T) and a metastatic carcinoma cell line (MDA-MB-231). The assay here does not distinguish primary proteins, which bind directly to ES rRNAs, from secondary proteins, which bind to primary proteins. Some of the proteins observed here in association with ESs are known to interact with each other in vivo suggesting secondary association.

Isolated ESs, or fragments of ESs, were attached to beads, incubated with cell lysates and washed to remove non-specific and weakly associated proteins. The remaining ES-associated proteins were proteolytically digested and the resulting peptides were identified by LC-MS/MS and analyzed by tandem mass spectra. ES-associated proteins were clustered based on biological process and/or molecular function with DAVID. Control experiments characterizing interactions of human proteins with an ES from a fungi (designated $ES7_{SC}$) were also performed.

For each rRNA fragment, four replica experiments were performed with each cell line. The proteins detected were partitioned into Groups (called Group I and II) based on the stringency of the result. Group I contains proteins observed in all four replica experiments, which are identified with the greatest confidence and appear to bind with the greatest affinity. Group II contains proteins observed in two or three of the four replica experiments, that appear to bind to the ESs with lower affinity or more transiently than Group I proteins. The groups associated with HEK 293T cells are indicated with a

subscript H (i.e. Group I_H). The groups associated with MDA-MB-231 cells are indicated with a subscript M (i.e. Group I_M). In general, more proteins were identified in Groups I_H than in Groups I_M.

A.3.2.2 Human ESs associate with the UPS

Many proteins of the ubiquitin-proteasome system (UPS) bind to human ESs. The amount of interaction depends on both the identity of the ES and on the cell line. Interactions with the UPS are more extensive for ES7 than for ES27, and are more extensive in HEK 293T cells than in MDA-MB-231 cells. The differences between cell lines are consistent with the abundance of UPS-proteins, which is known to depend on cell line and is higher in HEK 293T than in other cell lines tested.[5-6] Differences in protein affinity between ESs are not totally unexpected because ESs have distinctive sequences and structures.

The results here suggest that components of the UPS have specific affinity for ES7 and lower, non-specific affinity for general rRNAs. Sixty percent of proteins in Group $I_{ES7,H}$ (9 out of 15) and 12% in Group $I_{ES27,H}$ (2 out of 16) are known to be involved in ubiquitination or sumoylation processes. No Group $I_{ES7,M}$ and Group $I_{ES27,M}$ proteins are involved in the UPS. Group II proteins, which are thought to bind with low affinity, exhibit a more uniform distribution of UPS proteins over cell line and ES than that observed for Group I. UPS proteins constitute 6% of Group $II_{ES7,H}$ (12 out of 203), 5% of Group $II_{ES27,H}$ (12 out of 261), 7% of Group $II_{ES7,M}$ (12 out of 170) and 5% of Group $II_{ES27,M}$ (10 out of 182).

The results suggest that the UPS associates with ES7 in HEK 293T cells. Group $I_{ES7,H}$ contains a broad range of UPS proteins including enzymes involved in activation (E1), conjugation (E2) and ligation (E3) of ubiquitin, as well as known protein binding partners of ligase and deubiquitinating enzymes. Group $I_{ES7,H}$ contains three E1-like enzymes, SUMO-activating enzyme subunit 2 (SAE2), NEDD8-activating enzyme E1 regulatory subunit (ULA1) and Ubiquitin-like modifier-activating enzyme 5 (UBA5) and a E2 ubiquitin-conjugating enzyme, E2O (UBE2O). E2O is a hybrid of E2 and E3 activities. This Group also contains several proteins that bind to deubiquitinating enzymes, including transcription elongation factor A protein-like 4 (TCAL4), Ankyrin repeat domain-containing protein 17 (ANR17) and Nucleoside diphosphate kinase A (NME1).[7] CD2 antigen cytoplasmic tail-binding protein 2 (CD2B2) is a protein-binding partner for E3-ligase NEDD4 and Neurofilament light polypeptide (NFL) interacts with E3 SUMO-protein ligase PIAS4.[8-9] Group $I_{ES7,H}$ contains both, CD2B2 and NFL.

Unlike Group I, the proteins identified in Groups II are similar for both cell lines. Group $II_{ES7,H}$ include the 26S proteasome non-ATPase regulatory subunits 7 and 14, Ubiquitin carboxyl-terminal hydrolases 10 and 14, NEDD8-activating enzyme E1 catalytic subunit, Proteasome subunit beta type-7 and COP9 signalosome complex subunit 7a. Similarly, Group $II_{ES7,M}$ proteins include 26S proteasome non-ATPase regulatory subunit 8, Ubiquitin carboxyl-terminal hydrolase 4, NEDD8-activating enzyme E1 regulatory subunit, Proteasome subunit alpha type-2,4 and 6 and COP9 signalosome complex subunit 5.

Specific interactions of UPS components are attenuated in ES27 compared to ES7. No proteins in Group $I_{ES27,M}$ are involved in the UPS. Only two Group $I_{ES27,H}$

proteins are involved in the UPS. These two proteins are ubiquitin carboxyl-terminal hydrolase 7 (UBP7) and TIF1A, which has dual functionality and is involved in both transcription and protein degradation.[10-12] Group $II_{ES27,H}$ proteins include Ubiquitin-like modifier-activating enzyme 1, Ubiquitin-conjugating enzyme E2T, SUMO-activating enzyme subunit 1 and 26S proteasome non-ATPase regulatory subunit 10. Group $II_{ES27,M}$ proteins include Ubiquitin-like modifier-activating enzyme 6, Ubiquitin-conjugating enzyme E2C, SUMO-activating enzyme subunit 2 and NEDD8-activating enzyme E1 regulatory subunit.

Control experiments were performed with $ES7_{SC}$ (ES7 from *S. cerevisiae*) and HEK 293T cells. $ES7_{SC}$ differs from human ES7 in sequence and composition. $ES7_{SC}$ is smaller than ES7 in human although structurally the branching core is conserved. For $ES7_{SC}$, only 14% of proteins in Group I (1 out of 7) and 8% in Group II (14 out of 169) are related to the UPS. These results are in contrast with those obtained for human ES7 (Group $I_{ES7,H}$) in which 60% of the proteins are related to the UPS. In previous work, UPS components in *S. cerevisiae* cell extracts were not observed to associate with $ES7_{SC}$.[13]

A.3.2.3 Ribosomal proteins

Ribosomal proteins have at least some non-specific affinity for rRNA. However, ribosomal proteins that are not contained in assembled ribosomes are quickly degraded and are present at low levels in the cytosol.[14-15] Thus, we anticipated that even though ribosomal proteins are highly abundant in cells, they would not be accessible, and would not be identified in our experiments. Indeed, we observe only a single ribosomal protein

with high confidence. Ribosomal protein SA is observed in Group $I_{ES7,M}$. SA has multiple functions and is known to be present outside of ribosomes in the cytosol, in the plasma membrane and the in nucleus.[16] Except for L35, which is seen in both Group $II_{ES7,H}$ and Group $II_{ES27,H}$, there is no overlap between ribosomal proteins in various subsets of Group II. Group $II_{ES7,H}$ contains ribosomal proteins L6, L12, L35, L38, S4, S14, S21, S23 and S25. Group $II_{ES27,H}$ contains L7, L17, L31, L35, S2, S12 and S24. Group $II_{ES7,M}$ contains L13a, L28, L36a and S19. Group $II_{ES27,M}$ contains 60S acidic ribosomal protein P0, S6, S8, S24, S17 and S23. L6, L13a and L28 are among those that associate with ESs in the assembled ribosome.[17]

A.3.2.4 G-quadruplex associated proteins

Several proteins with known affinity for RNA G-quadruplexes were observed in Groups II. These proteins include RNA binding protein 14, ribosomal proteins L6, L7, L12, S2 and S6, and serine/arginine-rich splicing factors 1 and 9.[18] In addition, Fragile X mental retardation protein 1 (FMR1) and Fragile X mental retardation syndrome-related protein 1 (FXR1), which associates directly with FMR1, were also identified.[19-21]

A.3.2.5 Other ES7-associated proteins

Group $I_{ES7,H}$ contains proteins involved in transcription and translation. These proteins include TCAL4 and Eukaryotic peptide chain release factor subunit 1 (ERF1). TCAL4 was discussed in the section above as a binding partner for proteins associated with ubiquitination. Group $I_{ES7,H}$ contains a kinase, an ATPase, a nucleoside catabolic enzyme, a protein involved in RNA splicing and a structural protein. Group $I_{ES7,M}$ proteins differ from those of Group $I_{ES7,H}$ in that Group $I_{ES7,M}$ includes more proteins

involved in RNA processing. There are four proteins involved in mRNA processing within Group $I_{ES7,M}$ and one involved in transcription regulation.

For ES7, the functional classes of proteins are generally similar in Group I. Both Group $II_{ES7,H}$ and Group $II_{ES7,M}$ contain proteins involved in transcription, translation, mRNA splicing, signaling and structure. Group $II_{ES7,H}$, in comparison to Group $II_{ES7,M}$, contains more translation initiation factors and signaling proteins. Group $II_{ES7,H}$ contains some proteins that are not observed in Group $II_{ES7,M}$, including seryl-aminoacyl tRNA synthetase and several proteins involved in ribosomal biogenesis. Association of aminoacyl tRNA synthetases with ES7 was previously observed in *S. cerevisiae*.[13] Some Group II proteins may reflect weak secondary interactions between other Group II proteins. An example is given by eukaryotic initiation factor 4E (EIF4E) and Fragile X mental retardation protein (FMR1), found in Group $II_{ES7,H}$. A complex formed by FMR1, EIF4E and additional proteins inhibits translation initiation in neurons.[22] These observations reinforces that our experiments identify primary as well as secondary proteins.

A.3.2.6 Clustering of ES7-associated proteins

Clustering of ES-associated proteins was performed based on molecular function or biological process using DAVID (Figure A.2). Group $I_{ES7,H}$ shows four biological process clusters. These clusters are (i) protein catabolic processes, (ii) metabolic processes, (iii) ribonucleoside triphosphate biosynthetic processes and (iv) cellular response to stress. Clustering by molecular function gives only a single cluster, which is ligase.

Group $I_{ES7,M}$ shows three biological process clusters, (i) RNA processing, (ii) gene expression and (iii) cellular process. Clustering Group $I_{ES7,M}$ by molecular function gives only a single cluster, which is protein binding. As described previously, if an identical group of proteins is assigned to multiple clusters, we consider only the cluster with the lowest *P*-value.[13]

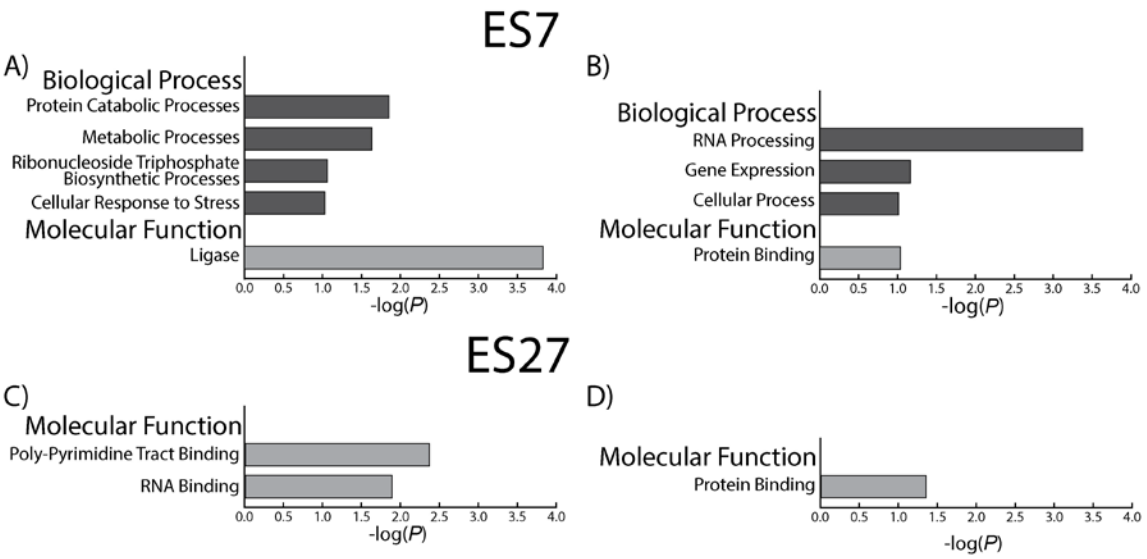

Figure A.2. Clustering of human ES7- and ES27-associated proteins obtained from pull-down experiments with HEK 293T and MDA-MB-231 cells based on biological process and molecular function. Clustering of proteins of (A) Group $I_{ES7,H}$, (B) Group $I_{ES7,M}$, (C) Group $I_{ES27,H}$, and (D) Group $I_{ES27,M}$. Clusters by biological process were not obtained for Group $I_{ES27,H}$ and Group $I_{ES27,M}$.

A.3.2.7 Other ES27-associated proteins

As with ES7, ES27 rRNA associates with proteins involved in RNA processing, transcription, translation and structure. Group $I_{ES27,H}$ contains many RNA binding

proteins and structural proteins. Group $I_{ES27,M}$ also include proteins involved in transcription, translation and RNA splicing.

As with ES7, in ES27 the distribution of Groups II proteins is generally similar to the distribution observed for Group I proteins. In fact, many protein categories identified here for ES27 were also identified in Groups II for ES7. Both, Group $II_{ES27,H}$ and Group $II_{ES27,M}$, contain proteins involved in transcription, translation, structure and signaling. Group $II_{ES27,H}$ contains a significant amount of structural proteins, several ATPases, GTPases, kinases, transcription factors, translation initiation factors, translation elongation factors, some splicing factors, a couple of ribosome biogenesis proteins, aminoacyl tRNA synthatases and RNA helicases. Although the composition of Group $II_{ES27,M}$ is similar to that of Group $II_{ES27,H}$, this group contains more translation initiation factors and aminoacyl tRNA synthetases, and lacks of proteins directly involved in RNA splicing, ribosome biogenesis, translation elongation factors and RNA helicases.

A.3.2.8 Clustering of ES27-associated proteins

Group $I_{ES27,H}$ and Group $I_{ES27,M}$ (Figure A.2C and D) failed to cluster by biological process. Molecular function of HEK 293T proteins yielded two clusters, (i) poly-pyrimidine tract binding and (ii) RNA binding. Clustering of MDA-MB-231 proteins yielded a single group, which is protein binding.

A.4 A universal chemical enrichment method for mapping the yeast N-glycoproteome by mass spectrometry

Glycosylation is one of the most common and important protein modifications in biological systems. Many glycoproteins naturally occur at low abundances, which makes comprehensive analysis extremely difficult. Additionally, glycans are highly heterogeneous, which further complicates analysis in complex samples. Lectin enrichment has been commonly used, but each lectin is inherently specific to one or several carbohydrates, and thus no single or collection of lectin(s) can bind to all glycans. Here we have employed a boronic acid-based chemical method to universally enrich glycopeptides. The reaction between boronic acids and sugars has been extensively investigated, and it is well known that the interaction between boronic acid and diols is one of the strongest reversible covalent bond interactions in an aqueous environment. This strong covalent interaction provides a great opportunity to catch glycopeptides and glycoproteins by boronic acid, whereas the reversible property allows their release without side effects. More importantly, the boronic acid-diol recognition is universal, which provides great capability and potential for comprehensively mapping glycosylation sites in complex biological samples.

By combining boronic acid enrichment with PNGase F treatment in heavy-oxygen water and MS, we have identified 816 N-glycosylation sites in 332 yeast proteins, among which 675 sites were well-localized with greater than 99% confidence. The results demonstrated that the boronic acid-based chemical method can effectively enrich glycopeptides for comprehensive analysis of protein glycosylation. A general trend seen within the large data set was that there were fewer glycosylation sites toward the C termini of proteins. Of the 332 glycoproteins identified in yeast, 194 were membrane proteins. Many proteins get glycosylated in the high-mannose N-glycan biosynthetic and GPI anchor biosynthetic pathways. Compared with lectin enrichment, the current method is more cost-efficient, generic, and effective. This method can be extensively applied to different complex samples for the comprehensive analysis of protein glycosylation.

A.5 Comprehensive analysis of protein N-glycosylation sites by combining chemical deglycosylation with LC–MS

Glycosylation is one of the most important protein modifications in biological systems. It plays a critical role in protein folding, trafficking, and stability as well as cellular events such as immune response and cell-to-cell communication. Aberrant protein glycosylation is correlated with several diseases including diabetes, cancer, and infectious diseases. The heterogeneity of glycans makes comprehensive identification of protein glycosylation sites very difficult by MS because it is challenging to match mass

spectra to peptides that contain different types of unknown glycans. We combined a chemical deglycosylation method with LC–MS-based proteomics techniques to comprehensively identify protein N-glycosylation sites in yeast. On the basis of the differences in chemical properties between the amide bond of the N-linkage and the glycosidic bond of the O-linkage of sugars, O-linked sugars were removed and only the innermost N-linked GlcNAc remained, which served as a mass tag for MS analysis. This chemical deglycosylation method allowed for the identification of 555 protein N-glycosylation sites in yeast by LC–MS, which is 46% more than those obtained from the parallel experiments using the Endo H cleavage method. A total of 250 glycoproteins were identified, including 184 membrane proteins. This method can be extensively used for other biological samples.

A.6 Systematic and site-specific analysis of N-sialoglycosylated proteins on the cell surface by integrating click chemistry and MS-based proteomics

Glycoproteins on the cell surface are ubiquitous and essential for cells to interact with the extracellular matrix, communicate with other cells, and respond to environmental cues. Although surface sialoglycoproteins can dramatically impact cell properties and represent different cellular statuses, global and site-specific analysis of sialoglycoproteins only on the cell surface is extraordinarily challenging. An effective

method integrating metabolic labeling, copper-free click chemistry and mass spectrometry-based proteomics was developed to globally and site-specifically analyze surface *N*-sialoglycoproteins. Surface sialoglycoproteins metabolically labeled with a functional group were specifically tagged through copper-free click chemistry, which is ideal because it is quick, specific and occurs under physiological conditions. Sequentially tagged sialoglycoproteins were enriched for site-specific identification by mass spectrometry. Systematic and quantitative analysis of the surface *N*-sialoglycoproteome in cancer cells with distinctive invasiveness demonstrated many *N*-sialoglycoproteins up-regulated in invasive cells, the majority of which contained cell adhesion-related domains. This method is very effective to globally and site-specifically analyze *N*-sialoglycoproteins on the cell surface, and will have extensive applications in the biological and biomedical research communities. Site-specific information regarding surface sialoglycoproteins can serve as biomarkers for disease detection, targets for vaccine development and drug treatment.

A.7 Systematic investigation of cellular response and pleiotropic effects in atorvastatin-treated liver cells by MS-based proteomics

For decades, statins have been widely used to lower cholesterol levels by inhibiting the enzyme HMG Co-A reductase (HMGCR). It is well-known that statins

have pleiotropic effects including improving endothelial function and inhibiting vascular inflammation and oxidation. However, the cellular responses to statins and corresponding pleiotropic effects are largely unknown at the proteome level. Emerging mass spectrometry-based proteomics provides a unique opportunity to systemically investigate protein and phosphoprotein abundance changes as a result of statin treatment. Many lipid-related protein abundances were increased in HepG2 cells treated by atorvastatin, including HMGCR, FDFT, SQLE, and LDLR, while the abundances of proteins involved in cellular response to stress and apoptosis were decreased. Comprehensive analysis of protein phosphorylation demonstrated that several basic motifs were enriched among down-regulated phosphorylation sites, which indicates that kinases with preference for these motifs, such as protein kinase A and protein kinase C, have attenuated activities. Phosphopeptides on a group of G-protein modulators were up-regulated, which strongly suggests that cell signal rewiring was a result of the effect of protein lipidation by the statin. This work provides a global view of liver cell responses to atorvastatin at the proteome and phosphoproteome levels, which provides insight into the pleiotropic effects of statins.

A.8 Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells

Protein dynamics are essential in regulating nearly every cellular event, and aberrant proteostasis is the source of many diseases. It is extraordinarily difficult to globally study protein dynamics and accurately measure their half-lives. Here we have developed a chemical proteomics method integrating protein labeling, click chemistry and multiplexed proteomics, which overcomes current challenges with existing methods. Labeling with both azidohomoalanine (AHA) and heavy lysine allows us to selectively enrich newly synthesized proteins, clearly distinguish them from existing proteins, and reduce the impact of heavy amino acid recycling. Moreover, multiplexed proteomics enables us to quantify proteins at multiple time points simultaneously, thus increasing the accuracy of measuring protein abundance changes and their half-lives. Systematic investigation of newly synthesized protein dynamics will provide insight into proteostasis and the molecular mechanisms of disease.

A.9 Quantification of tunicamycin-induced protein expression and N-glycosylation changes in yeast

Xiao, H., Smeekens, J.M., Wu, R. Quantification of tunicamycin-induced protein expression and N-glycosylation changes in yeast, *Analyst*, 2016, 141, 3737-3745.

Tunicamycin is a potent protein N-glycosylation inhibitor that has frequently been used to manipulate protein glycosylation in cells. However, protein expression and glycosylation changes as a result of tunicamycin treatment are still unclear. Using yeast as a model system, we systematically investigated the cellular response to tunicamycin at the proteome and N-glycoproteome levels. By utilizing modern mass spectrometry-based

proteomics, we quantified 4259 proteins, which nearly covers the entire yeast proteome. After the three-hour tunicamycin treatment, more than 5% of proteins were down-regulated by at least 2 fold, among which proteins related to several glycan metabolism and glycolysis-related pathways were highly enriched. Furthermore, several proteins in the canonical unfolded protein response pathway were up-regulated because the inhibition of protein N-glycosylation impacts protein folding and trafficking. We also comprehensively quantified protein glycosylation changes in tunicamycin-treated cells, and more than one third of quantified unique glycopeptides (168 of 465 peptides) were down-regulated. Proteins containing down-regulated glycopeptides were related to glycosylation, glycoprotein metabolic processes, carbohydrate processes, and cell wall organization according to gene ontology clustering. The current results provide the first global view of the cellular response to tunicamycin at the proteome and glycoproteome levels.

A.10 References

1. Chung, Y. I.; Ahn, K. M.; Jeon, S. H.; Lee, S. Y.; Lee, J. H.; Tae, G., Enhanced bone regeneration with BMP-2 loaded functional nanoparticle-hydrogel complex. *J Control Release* **2007,** *121* (1-2), 91-99.

2. Fleischer, T. C.; Weaver, C. M.; McAfee, K. J.; Jennings, J. L.; Link, A. J., Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. *Gene Dev* **2006,** *20* (10), 1294-1307.

3. Beckmann, B. M.; Horos, R.; Fischer, B.; Castello, A.; Eichelbaum, K.; Alleaume, A. M.; Schwarzl, T.; Curk, T.; Foehr, S.; Huber, W.; Krijgsveld, J.; Hentze, M. W., The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **2015,** *6*.

4. Huh, W. K.; Falvo, J. V.; Gerke, L. C.; Carroll, A. S.; Howson, R. W.; Weissman, J. S.; O'Shea, E. K., Global analysis of protein localization in budding yeast. *Nature* **2003,** *425* (6959), 686-691.

5. Kaiser, S. E.; Riley, B. E.; Shaler, T. A.; Trevino, R. S.; Becker, C. H.; Schulman, H.; Kopito, R. R., Protein standard absolute quantification (PSAQ) method for the measurement of cellular ubiquitin pools. *Nat. Methods* **2011,** *8* (8), 691-U129.

6. Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **2012,** *11* (3).

7. Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W., Defining the human deubiquitinating enzyme interaction landscape. *Cell* **2009,** *138* (2), 389-403.

8. Kofler, M.; Motzny, K.; Beyermann, M.; Freund, C., Novel interaction partners of the CD2BP2-GYF domain. *J Biol Chem* **2005,** *280* (39), 33397-33402.

9. Goehler, H.; Lalowski, M.; Stelzl, U.; Waelter, S.; Stroedicke, M.; Worm, U.; Droege, A.; Lindenberg, K. S.; Knoblich, M.; Haenig, C.; Herbst, M.; Suopanki, J.; Scherzinger, E.; Abraham, C.; Bauer, B.; Hasenbank, R.; Fritzsche, A.; Ludewig, A. H.; Buessow, K.; Coleman, S. H.; Gutekunst, C. A.; Landwehrmeyer, B. G.; Lehrach, H.; Wanker, E. E., A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell* **2004,** *15* (6), 853-865.

10. LeDouarin, B.; Nielsen, A. L.; Garnier, J. M.; Ichinose, H.; Jeanmougin, F.; Losson, R.; Chambon, P., A possible involvement of TIF1 alpha and TIF1 beta in the epigenetic control of transcription by nuclear receptors. *Embo J* **1996,** *15* (23), 6701-6715.

11. Le Douarin, B.; You, J.; Nielsen, A. L.; Chambon, P.; Losson, R., TIF1 alpha: A possible link between KRAB zinc finger proteins and nuclear receptors. *J Steroid Biochem* **1998,** *65* (1-6), 43-50.

12. Allton, K.; Jain, A. K.; Herz, H. M.; Tsai, W. W.; Jung, S. Y.; Qin, J.; Bergmann, A.; Johnson, R. L.; Barton, M. C., Trim24 targets endogenous p53 for degradation. *P Natl Acad Sci USA* **2009,** *106* (28), 11612-11616.

13. Ramos, L. M. G.; Smeekens, J. M.; Kovacs, N. A.; Bowman, J. C.; Wartell, R. M.; Wu, R. H.; Williams, L. D., Yeast rRNA expansion segments: Folding and function. *Journal of Molecular Biology* **2016,** *428* (20), 4048-4059.

14. Lam, Y. W.; Lamond, A. I.; Mann, M.; Andersen, J. S., Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. *Curr Biol* **2007,** *17* (9), 749-760.

15. Sung, M. K.; Reitsma, J. M.; Sweredoski, M. J.; Hess, S.; Deshaies, R. J., Ribosomal proteins produced in excess are degraded by the ubiquitin-proteasome system. *Mol Biol Cell* **2016,** *27* (17), 2642-2652.

16. DiGiacomo, V.; Meruelo, D., Looking into laminin receptor: critical discussion regarding the non-integrin 37/67-kDa laminin receptor/RPSA protein. *Biol Rev* **2016,** *91* (2), 288-310.

17. Anger, A. M.; Armache, J. P.; Berninghausen, O.; Habeck, M.; Subklewe, M.; Wilson, D. N.; Beckmann, R., Structures of the human and Drosophila 80S ribosome. *Nature* **2013,** *497* (7447), 80-+.

18. von Hacht, A.; Seifert, O.; Menger, M.; Schutze, T.; Arora, A.; Konthur, Z.; Neubauer, P.; Wagner, A.; Weise, C.; Kurreck, J., Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic Acids Res* **2014,** *42* (10), 6630-6644.

19. Darnell, J. C.; Jensen, K. B.; Jin, P.; Brown, V.; Warren, S. T.; Darnell, R. B., Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* **2001,** *107* (4), 489-499.

20. Zanotti, K. J.; Lackey, P. E.; Evans, G. L.; Mihailescu, M. R., Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry-Us* **2006,** *45* (27), 8319-8330.

21. Siomi, M. C.; Zhang, Y.; Siomi, H.; Dreyfuss, G., Specific sequences in the fragile X syndrome protein FMR1 and the FXR proteins mediate their binding to 60S ribosomal subunits and the interactions among them. *Mol Cell Biol* **1996,** *16* (7), 3825-3832.

22. Napoli, I.; Mercaldo, V.; Boyl, P. P.; Eleuteri, B.; Zalfa, F.; De Rubeis, S.; Di Marino, D.; Mohr, E.; Massimi, M.; Falconi, M.; Witke, W.; Costa-Mattioli, M.;

Sonenberg, N.; Achsel, T.; Bagni, C., The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell* **2008,** *134* (6), 1042-1054.