Chemical Biology Thesis Background Exam

Elucidating the Membrane Proteoform-ome

March 13, 2019

Project Overview

Membrane proteins play crucial roles in cellular function and make up over a quarter of the human genome.[1–6] The hydrophobic regions of integral membrane proteins that allow for association with the plasma membrane also induce protein insolubility in aqueous solutions. Insolubility in conjunction with low expression levels both contribute to the underrepresentation of membrane proteins in whole proteome datasets.[1,7] Progress has been made in enhancing the identification of membrane proteins with bottom-up proteomics[8–12], but these methods do not provide proteoform level information. Proteoforms are the different forms of proteins produced from the same gene due to genetic variations, RNA splice variants, and post-translational modifications.[13] Proteoform identification is essential for understanding any biological system because proteoforms are the acting molecules in a vast array of cellular functions.[14–16] In top-down and intact mass proteomics, individual proteoforms are identified with mass spectrometry.[17–21] Membrane proteoforms have been studied in the past utilizing top-down and intact mass strategies[22–25], but obstacles such as low abundance and insolubility plague comprehensive proteoform analysis. The goal of this project is to develop and enhance strategies for the identification of membrane proteoforms. Our first aim is to combat membrane protein insolubility and low abundance through the development and improvement of membrane proteoform enrichment and solubilization strategies. Enhanced membrane proteoform enrichment will be accomplished through the synthesis of a novel, biotinylated cell surface capture tag. Proteoform solubility in various solution conditions will be optimized through the use of organic solvents, organic acids and mass spectrometric compatible detergents. Our second aim is to enhance the sensitivity of mass spectral analysis of membrane proteins to increase the number of proteoforms identified. Separation of membrane proteoforms prior to mass spectral analysis will be optimized to prevent complications with proteoform insolubility. Implementation of various instrument control strategies will enhance the identification of low abundance species by intelligent ion collection and minimizing repetitive proteoform fragmentation. Finally, the developed strategies will be applied to elucidate the membrane proteoforms of pluripotent stem cells. This research will facilitate the large-scale study of membrane proteoforms, developing a deeper knowledge of membrane proteins than has previously been possible.

Chemical Biology Thesis Background Exam

Elucidating the Membrane Proteoform-ome

March 13, 2019

Contents

Project Overview 1

Table of Contents 2

1 Curriculum Vitae 3

2 Research Plan 6

2.1 Specific Aims……6

2.2 Background……7

2.3 Approach……8

References 17

Specific Aims

Over a quarter of the human genome encodes for membrane proteins[2,3]. These proteins play critical roles in cellular function, making them valuable drug targets for disease intervention. Even though their importance is acknowledged, membrane proteins are grossly understudied compared to their intercellular protein counterparts due in part to insolubility and low abundance.[1,7] Proteomic analysis has the potential to provide a global understanding of membrane proteins that no other biochemical technique can rival.[26] The use of bottom-up proteomics, which utilizes identified peptides to infer which protein are present in a sample, has made strides in the analysis of membrane proteins. Valuable proteoform information is lost through the digestion process, making bottom-up proteomics incapable of proteoform level analysis.[7] Whole membrane proteins analyzed with top-down and intact mass proteomic strategies can provide knowledge of what proteoforms are present in a dynamic biological system. This information is indispensable to the development of a complete understanding of the complex phenotypes and functional regulation afforded by membrane proteoforms.

Although proteoform analysis of membrane proteins will provide a crucial understanding of these under-studied work-horses, there are still challenges regarding their analysis that must be overcome before any large-scale studies can be successful. Their insolubility and low abundance is at odds with proteoform analysis which favors high abundance, low molecular weight and soluble proteins. We propose here to expand the depth to which membrane proteoforms are characterized by developing technologies for improved enrichment, solubility and separation as well as enhancing the sensitivity of mass spectrometric analysis of low abundance proteoforms. The proposed work is summarized in the following aims:

Aim 1: Optimize and develop strategies for membrane proteoform enrichment and solubilization.

- Adapt cell surface capture technology for the elucidation of the membrane proteoforms.
- Enhance membrane protein solubilization conditions by evaluating the use of MS-comatible surfactants, organic solvents and organic acids.

Aim 2: Improve mass spectral analysis of membrane proteins to increase sensitivity and proteoform identifications.

- Improve online separation of membrane proteins to allow for efficient delivery to mass spectrometer.
- Advance and implement real-time instrument control to expand proteoform identifications by providing guided selection of MS candidates.
- Develop informatics strategies incorporating multiple data types for improved proteomic analysis and creation of sample-specific databases.

Aim 3: Elucidate the membrane proteoforms of pluripotent stem cells.

- Generate a proteogenomic, multi-protease sample-specific database.
- Perform cell surface proteoform enrichment for top-down and intact mass analysis.

Background and Significance

Membrane proteins operate as transporters, channels, receptors and enzymes that are responsible for cell to cell communication and for the exchange of information between cells and their environment. [4–6] Approximately 20-30% of the human genome encodes for membrane proteins[2,3], but compared to their soluble, intercellular protein counterparts, very little is known about them.[7] The transmembrane region of membrane proteins contains hydrophobic amino acids, while the extracellular and intracellular domains contain more hydrophilic residues. Membrane proteins are not soluble in the aqueous environment necessary for their biochemical study.[1] When removed from the membrane, these proteins easily precipitate and form hydrophobic aggregates.[1] Membrane proteins also represent over 50% of current of drug targets.[27] This number will likely continue to increase with the growing use of antibodies to target surface proteins involved in tumor immune evasion.[28] Developing a better understanding of the proteins residing in and on the cellular membrane would be invaluable in developing effective and targeted medications.

Proteomic analysis utilizing mass spectrometry (MS) elucidates information about the proteins within a system such as their identity, abundance and any modifications present. In bottom-up proteomics, proteins are enzymatically digested into peptides that are analyzed via mass spectrometry.[29,30] Although proteins can be identified in bottom-up experiments, valuable information about the proteoforms present in a sample are lost. Proteoforms are all the various molecular forms from a single gene. Proteoforms can result from genetic mutations, alternatively spliced mRNA transcripts, and post-translational modifications.[13] Many proteoforms of a single gene or from different genes can share peptides, making bottom-up proteomics incapable of providing proteoform level identifications. Knowledge of the proteoforms present in a biological system is critical in developing an understanding of all the intricacies that contribute to the complex phenotypes we observe.[14–16] Top-down and intact mass proteomics analyzes the whole protein allowing for proteoforms to be both identified and quantified.[17–21]

Proteomic analyses have the ability to provide a global understanding of membrane proteoforms that is not achievable by other biochemical techniques.[26] However, there are still many barriers to success that must be overcome before a comprehensive catalogue of membrane proteoforms can be created. Proteomic analyses generally favor soluble and abundant proteins.[31] Membrane proteins must be solubilized in mass spectrometric compatible conditions, during sample preparation and chromatographic separation, in order to be analyzed via mass spectrometry. Membrane proteins are low abundance and their signals are overwhelmed by that of high abundance soluble proteins. These high abundance proteins are then preferentially selected for fragmentation.[31] This effect is exaggerated in top-down proteoform analysis that is 10-20 fold less sensitive then bottom-up proteomics due to a large number of charge states and isotopomers.[18] Enrichment of membrane proteins from complex cell lysate must occur to offset their low abundance and allow for fragmentation. The sensitivity of mass spectrometric analysis must also be enhanced to prevent repetitive fragmentation of a few highly abundant proteoforms. Surpassing these obstacles will facilitate the ability for the creation of membrane proteoform catalogues for various cell and tissue types, increasing the scientific communities understanding of membrane proteins.

Approach

Aim 1: Optimize and develop strategies for membrane proteoform enrichment and solubilization.

Adapt cell surface capture technology for the elucidation of the cell membrane proteoforms. Biotinylated tags targeting functional groups on extracellular domains of membrane proteins can be used to enrich these proteins from complex cell lysate. This strategy has been utilized successfully in bottom-up proteomics to gain information about the cell surface proteome. [32,12,11,33–36] The general steps of this process include tagging of cell surface proteins on live cells, cell lysis, digestion of proteins, and enrichment of the tagged peptides leveraging biotin's affinity for streptavidin. We propose the adaptation of this strategy for top-down and intact mass proteoform analysis.

Two main tags have been utilized for biotin-enrichment cell surface capture. The first tag is sulfo-NHS-SS-biotin (Figure 1A).[37] This tag reacts with primary amines such as extra cellular lysine residues. [38] A benefit of this tag is its cleavability allowing for easy removal. Unfortunately, multiple factors contribute to this tag having a low capture rate for membrane proteins such as the fact that not all membrane proteins have accessible extracellular lysine residues and that the hydrolysis of the NHS-ester competes with the tagging reaction.[34] The second tag, biocytin hydrazide (Figure 1B), is considered to be the gold standard of cell surface capture. It reacts with oxidized sugar residues on cell surface glycoproteins to form an extremely stable Schiff's base called a hydrazone. [39] This tag has a much higher membrane protein capture rate than sulfo-NHS-SS-biotin[32] but can only be removed by the cleavage of the glycosylation with PNGaseF. This is undesirable because valuable glycoform information is lost. We propose the synthesis of biotin-SS-hydrazide (Figure 2), a novel, cleavable tag targeting glycosylated cell surface proteins which combines the positive aspects of both tags without any of the negatives. The proposed synthetic route to this molecule can be observed in Figure 2. Briefly, an amide bond will be formed combining biotin and cysteamine utilizing dicyclohexylcarbodiimide to generate a thiol intermediate.[40] Next, a disulfide bond will be generated between the thiol intermediate and methyl 3-mercaptopropionate to form a biotin disulfide linked to a methyl ester.[41] Hydrazine will react with the methyl ester to generate biotin-SS-hydrazide, the desired end product.[42]

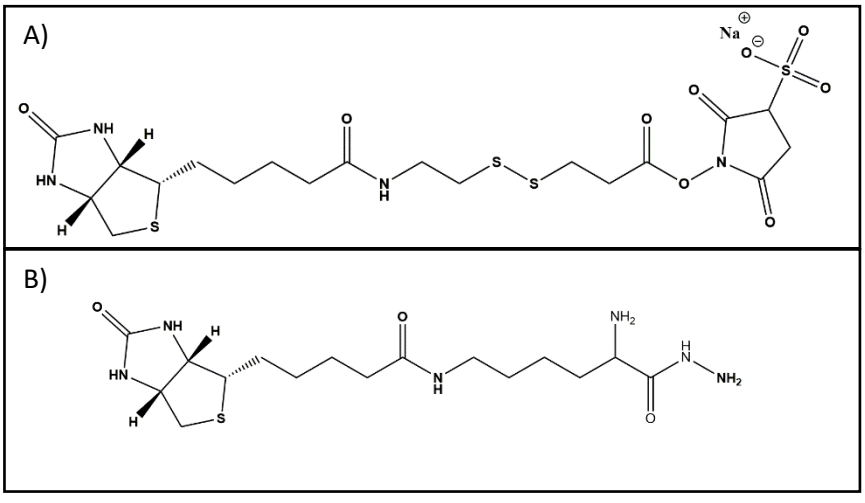


Figure 1: Chemical structures of A) Sulfo-NHS-SS-Biotin and B) Biocytin Hydrazide.

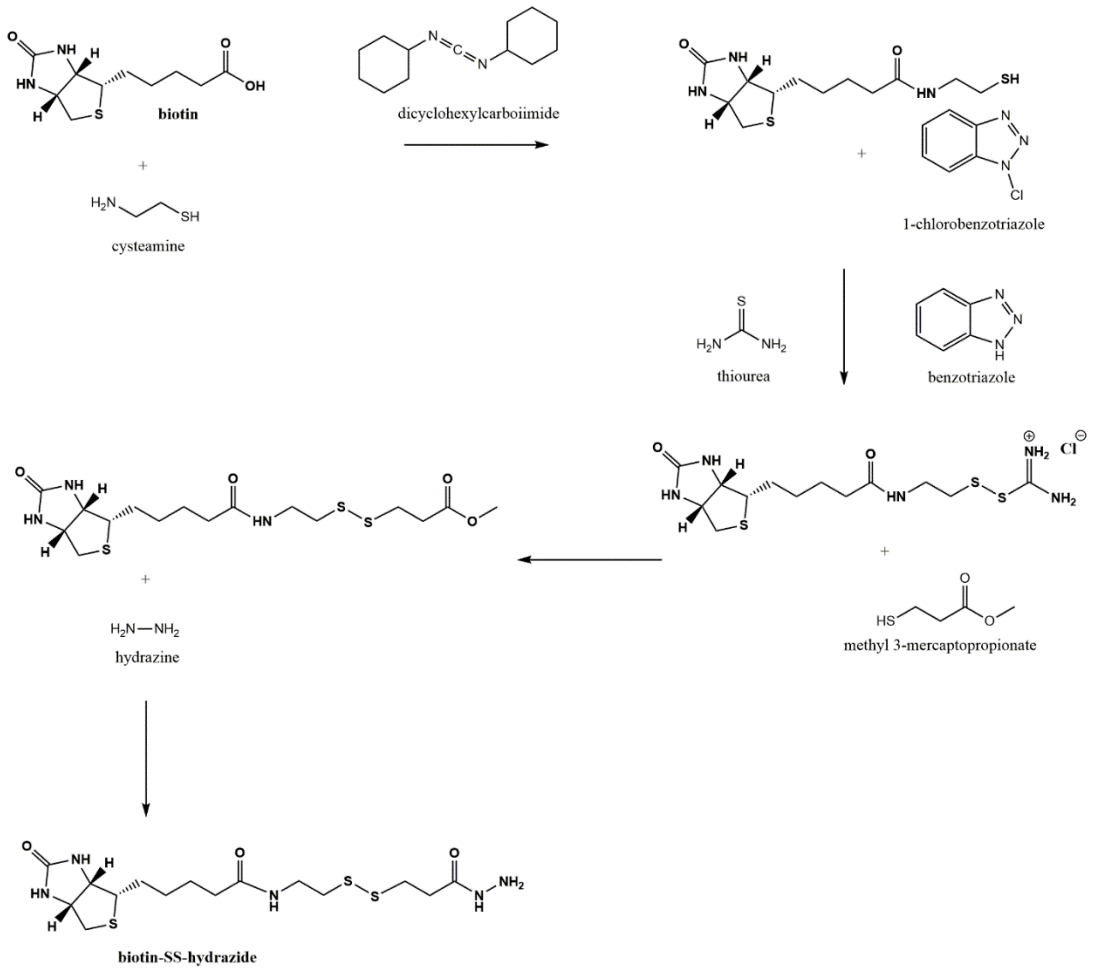


Figure 2: Synthetic scheme for the generation of Biotin-SS-Hydrazide.

The tagging, enrichment and cleavage efficiency of biotin-SS-hydrazide will be evaluated utilizing a mixture of a glycosylated antibody protein standard (SILu Lite SigmaMab (Sigma-Aldrich)) and BSA acting as a negative control. The heavy chain of SILu Lite SigmaMab has a well characterized glycosylation site and 3 documented glycoforms. Biotin-SS-hydrazide will be added to a solution containing both BSA and SILu Lite SigmaMab. Selectivity of the tagging reagent will be evaluated by ensuring that only SigmaMab is present after enrichment. Tagging efficiency can be evaluated by determining the number of tagging sites present on the various glycoforms and then evaluating what fraction of these sites were tagged by localizing the mass shift of 119 Da corresponding to the cleaved tag. Finally, we will evaluate the cleavage efficiency of the tag by comparing the known concentration of SILu Lite SigmaMab in the mixture to the retrieved concentration after capture, cleavage and elution. Protein concentration will be monitored by Nano-drop. Any loss in protein concentration would correspond to the linker not being cleaved and failing to elute.

Following evaluation of biotin-SS-hydrazide in the reduced complexity sample, we propose comparison to the gold standard biocytin hydrazide. Bottom-up and top-down experiments will be conducted on live Jeko cells. Two aliquots of cells will be utilized for each experiment, one will be

labeled with biocytin hydrazide and the other with our biotin-SS-hydrazide. For the bottom-up study, aliquots will be oxidized, labeled with their tag, lysed, and digested. The glycopeptides will be enriched with a streptavidin column and then glycosylations will be removed with PNGaseF.[12,11] The resulting peptides identified between the two methods will be compared. Proteoform identification will be compared after utilizing top-down proteomics in combination with PNGaseF to remove the glycosylations after capture with either biotin hydrazide or biotin-SS-hydrazide.

Alternative path to success. If biotin-SS-hydrazide fails to have good cleavage efficiency or if the synthesis is unsuccessful, we propose the synthesis of desthiobiotin hydrazide following the original biocytin hydrazide synthesis protocol.[42] The use of desthiobiotin allows for elution of proteoforms off the streptavidin column with biotin. Failure of the desthiobiotin hydrazide tag would require the use of biocytin hydrazide and PNGaseF to analyze proteoforms without glycoform analysis. We propose the isolation of membrane proteins utilizing differential and density gradient centrifugation as an alternative to biotin enrichment tags. This technique is well-developed and is frequently utilized in the proteomic community for the study of membrane proteins [43–46] but has a higher level of non-membrane protein contaminants compared to cell surface capture approaches.[47]

Enhance membrane protein solubilization conditions by evaluating the use of MS-compatible surfactants, organic solvents and organic acids. Membrane proteins tend to precipitate and form hydrophobic aggregates in the aqueous solvent conditions conventionally utilized for mass spectrometric analysis.[1] Surfactants, organic solvents and acids have been used to solubilize membrane proteins for proteomic analysis. We propose the evaluation of various MS-compatible surfactants (Rapigest[48] and Invitrosol), organic solvents (methanol, ethanol, 2-propanol, isopropanol and trifluoroethanol)[49–53] and organic acids (formic acid and trifluoroacetic acid)[54–56] to solubilize membrane proteins utilizing a Nano-drop derived assay. Membrane proteins will be enriched from Jeko cell lysate and divided into aliquots one for each solubilization technique and one for 95:5 water: acetonitrile 0.1% FA, a conventional aqueous solvent system for MS analysis. Each solvent condition will be utilized to solubilize membrane proteins within the sample. Samples will then be spun down to remove any precipitate and supernatant will be analyzed with Nano-drop taking absorbance values at 215 nm. The most effective solubilization condition will be determined by which technique provides the greatest absorbance by Nano-drop and will be utilized for all membrane proteoform analyses going forward.

Aim 2: Improve mass spectral analysis of membrane proteins to increase sensitivity and proteoform identifications.

Improve online separation of membrane proteoforms to allow for efficient delivery to mass spectrometer. Membrane proteins can cause complications with chromatographic separations that typically precede mass spectrometric analysis due to their increased hydrophobicity.[57] They can have strong hydrophobic interactions with the column's walls or with the stationary phase and can even precipitate directly onto the column when the mobile phase becomes too hydrophilic.[1,57] We propose optimization of the mobile and stationary phases utilized in chromatographic separation of membrane proteins to improve delivery to the mass spectrometer.

We propose the use of hydrophilic liquid chromatography (HILIC) as an online separation technique for membrane proteoforms. HILIC utilizes hydrophilic stationary phases with reverse-phase type eluents.[58,59] The typical composition of a HILIC mobile phase is mostly organic solvent with a small aqueous component. These solvent conditions are much more compatible with maintaining membrane

protein solubility. Although membrane proteins are considered to be hydrophobic in nature, the hydrophilic residues and glycans on membrane proteins will interact with the stationary phase to contribute to separation of proteoforms.[60] HILIC has been effectively used to fractionate mitochondrial membrane proteins prior to reverse-phase LC-MS analysis[58] and has in other cases been applied as a front end separation technique for intact protein analysis[59]. We propose the use of PolyHYDROXYETHYL A as the packing material for our HILIC columns. The organic solvent or mixture of solvents utilized for the mobile phase will be optimized for the separation of membrane proteoforms. The efficacy of HILIC as an online separation technique will be compared to conventional reverse phase liquid chromatography by analyzing aliquots of membrane proteins isolated by differential centrifugation. The total number of membrane proteoforms identified as well as the overlap of identifications between the two techniques will be compared.

We propose the use of CZE-MS/MS in addition to HILIC-MS/MS analysis to act as an orthogonal separation technique and increase the total number of membrane proteoforms identified. Capillary electrophoresis has been shown to be an effective method of separating proteoforms.[61] In the 2018 study by the Sun group, the ratio of membrane proteoforms to intercellular proteoforms identified by CZE-MS/MS was comparable to the ratio of membrane proteins to intercellular proteins within the UniProt database, signifying that their linear polyacrylamide capillary coating and buffer systems are at least somewhat compatible with membrane proteoform analysis.[61] Obtaining proteoform data from both CZE-MS/MS and HILIC-MS/MS will lead to the greatest number of proteoform identifications with the most sensitivity.

<u>Advance and implement real-time instrument control to expand proteoform identifications by providing guided selection of MS candidates.</u>

We propose implementation of real-time instrument control to improve the selection of candidates for fragmentation. We seek to alter the number of times an ion is selected for fragmentation and the manner in which MS1 spectra are obtained to improve the detection of low abundance species.

Improved MS1 Spectra with Enhanced Sensitivity. Proteoform ions selected for fragmentation are chosen from the set of ions observed in the intact mass spectrum (MS1). Low abundance species can be difficult to observe and are rarely selected for fragmentation because ions fill the C or ion - trap proportionally to their concentration. In 2018, Meier et al. introduced the boxcar technique in which ions are collected from small overlapping m/z windows for a time that is inversely proportional to the ion current from that range.[62] This technique enhances the signal for low abundance ions allowing them to be observed and selected for fragmentation. We propose the adaptation of the boxcar strategy for intact mass proteoform analysis and hypothesize that its use will improve the variety of proteoform identified.

Improved Precursor Selection. Conventional DDA experiments select the most abundant ions in each MS1 spectrum for fragmentation. The ion signal in top-down MS2 scans is spread out over more fragments meaning there are fewer ions of any given fragment and that more time is required to obtain sufficient signal to noise. Because of this increased time requirement, typically only the top 2 most abundant peaks are selected for fragmentation, compared to the top-10 or top-20 allowed by bottom-up proteomics. An ion is often excluded from fragmentation for a brief window after already being selected 2-3 times to prevent continual fragmentation of a single species. Many highly abundant ions continue to be selected for fragmentation even with this exclusion parameter, therefore, low

abundance ions are rarely selected for fragmentation. This bias is particularly prevalent in proteoform analysis where multiple charge states and isotope peaks are present in an MS1 spectrum providing more opportunities for a single proteoform to be selected for fragmentation multiple times.[18] We propose to implement real-time analysis[63] to increase the variety of peaks selected and to improve the rate of successful identification of both peptides and proteoforms. Our lab has already developed the open-source search software, MetaMorpheus, which is capable of both peptide and proteoform identification. This software program will be adapted for use in this project. We also have access to real-time control of our QE-HF orbitrap mass spectrometer through the Thermo instrument application programming interface (IAPI). We will implement code to analyze each tandem mass spectrum obtained by the instrument in real time. Peaks successfully identified on the first try will be excluded from reanalysis. Peaks not successfully identified will be subject to reanalysis possibly using a secondary charge state. We will also perform real-time deconvolution of MS1 spectra enabling us to discern which MS1 peaks are shared within a single proteoform to avoid reanalyzing the same peak multiple times.

Develop informatics strategies incorporating multiple-data types for improved proteomic analysis and creation of sample-specific databases. Samples for membrane proteoform analysis will be prepared for three separate analyses: RNA extraction for RNA-Seq analysis; multi-protease bottom-up analysis of enriched membrane proteins; and top-down/intact mass analysis of enriched membrane proteoforms. Peptides and proteoforms containing unknown sequence variants or unknown splice junctions cannot be identified by standard proteomic approaches because their sequences are not present in the conventional UniProt database.[64] We propose to address this complication utilizing a proteogenomic approach that couples deep transcriptomic and proteomic analyses. The obtained RNA-Seq data will be mined for variants that code for non-synonymous mutations in the proteome, as well as novel splice junctions, insertions and deletions, that are then coded into a protein sequence database. This process is accomplished utilizing Spritz, a software program developed in our lab that integrates more than 20 bioinformatics tools.

Multi-protease bottom-up proteomic analysis increases protein sequence coverage, reveals a wide variety of PTM-modified peptides present in the proteome, and eliminates superfluous protein entries from the resulting protein pruned database. Our lab has developed a novel bioinformatics strategy that allows for the comprehensive identification and localization of both known and unknown PTMs in bottom-up proteomics data without increasing the incidence of false positives.[65,66] The global PTM discovery (G-PTM-D) strategy also facilitates the incorporation of identified PTMs into a protein database for improved proteoform analysis.[65] This generated database is pruned, eliminating any protein entries that were not identified by bottom-up proteomics. Conventionally, trypsin is the protease of choice for bottom-up analyses, but the use of multiple proteolytic enzymes in parallel provides access to regions of the proteome that are not accessible when using only one protease.[67] This is especially true within transmembrane domains where lysine and arginine residues are less frequent. [68] Analysis of peptides from different proteolytic digestion leads to increased sequence coverage of analyzed proteins. Coupling this increased sequence coverage with G-PTM-D leads to an increase in the number of novel modifications that can be identified, localized and added to the protein database.[69] Additionally, the use of multiple proteases improves results of protein inference by disambiguating protein groups for more confident identifications leading to a more accurate protein pruned database.[69]

Preliminary Studies.

Integrated Top-down and Multi-Protease Bottom-up Proteomics of Jurkat.[69]

Aliquots of Jurkat cell lysate were digested by either Arg-C, Asp-N, chymotrypsin, Glu-C, Lys-C or trypsin and then subjected to high pH offline fractionation (10-11x) prior to LC-MS/MS analysis. Analysis of all of the multi-protease spectra with MetaMorpheus at once ("combined approach") showed an increase in the number of PTMs identified with G-PTM-D as well as an improvement in protein inference compared to separate proteolytic analyses ("separate approach") and compared to a single tryptic digest.

The combined multi-protease approach demonstrated an 18% increase in the number of single protein group identifications and an 8% decrease in the average number of protein accessions per protein group as compared to the separate multi-protease approach. The reduction in the number of proteins per group from the separate to the combined multi-protease approach is illustrated in Figure 3A. An entrapment study was also conducted showing the combined multi-protease approach led to a 33% decrease in the number of entrapped protein groups over the separate multi-protease approach. When compared to the tryptic digest alone, the combined multi-protease approach showed a 36% increase in the number of single protein group identifications and a 10% decrease in the average number of protein accessions per protein group. The reduction in the number of proteins per group from the tryptic digest to the combined multi-protease approach is illustrated in Figure 3B.

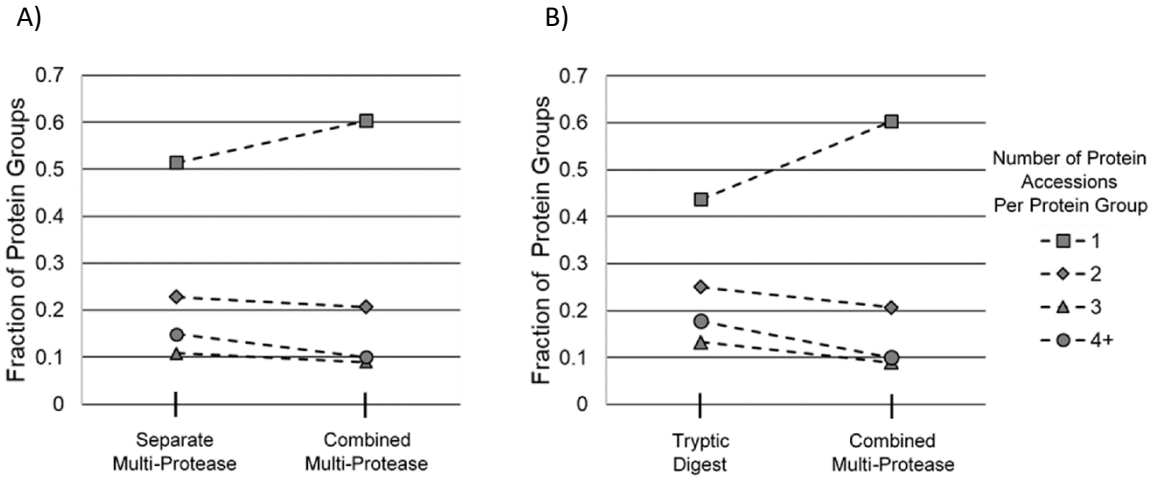


Figure 3: Plots of the fraction of all protein groups containing between 1 and 4+ protein Reduction in the ambiguity of the protein group identification can be observed when comparing the combined multi-protease approach with A) the separate multi-protease approach; or with B) the tryptic digest.

The use of multiple-proteases also increased the number of PTMs discovered by G-PTM-D. A total 6,007 PTMs of biological origin were identified and localized to a single amino acid residue. Approximately 75% of the modifications were unique to a single proteolytic digestion. The use of only

one protease resulted in a 3 to 17-fold decrease in the number of biologically relevant PTMs identified, depending on the protease (Figure 4).

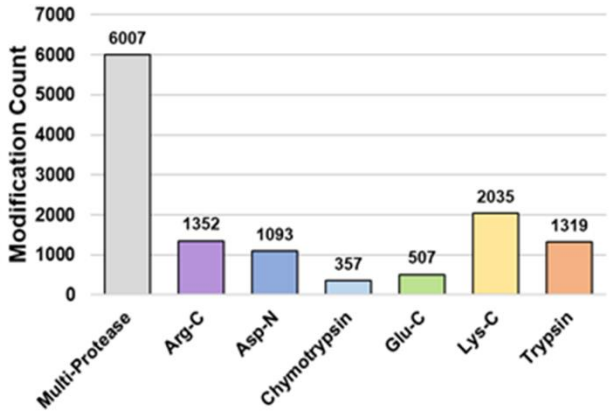


Figure 4: Bar graph comparing the number of modifications identified from the combined multi-protease approach to the number of modifications identified by peptide identifications from each of the six individual proteases.

The use of multi-protease protein pruned database for proteoform identifications. The use of a MetaMorpheus protein pruned database to generate a theoretical proteoform database within Proteoform Suite has been shown to increase the number of proteoform identifications while controlling the overall FDR.[70] We show an additional improvement in the number of proteoforms identified when utilizing a multi-protease protein pruned database. Intact mass Jurkat data was searched against the conventional UniProt protein, trypsin protein pruned, and multi-protease protein pruned databases. The inclusion of modifications identified by G-PTM-D increases the number of proteoforms identified while protein pruning helps keep FDR stable despite the increase in potential modified proteoforms within the database (Figure 5). The number of proteoforms identified using the UniProt database was 388. The use of trypsin generated protein pruned database yielded 439 proteoform identifications, a 13% increase in proteoform identifications. The use of the multi-protease protein pruned database produced 491 proteoform identifications, a 12% increase from the use of the trypsin database and a 27% increase over the use of the UniProt database. This increase in proteoforms can be attributed to the increase in modifications that are present due to G-PTM-D and due to the use of multi-protease to identify more PTMs.

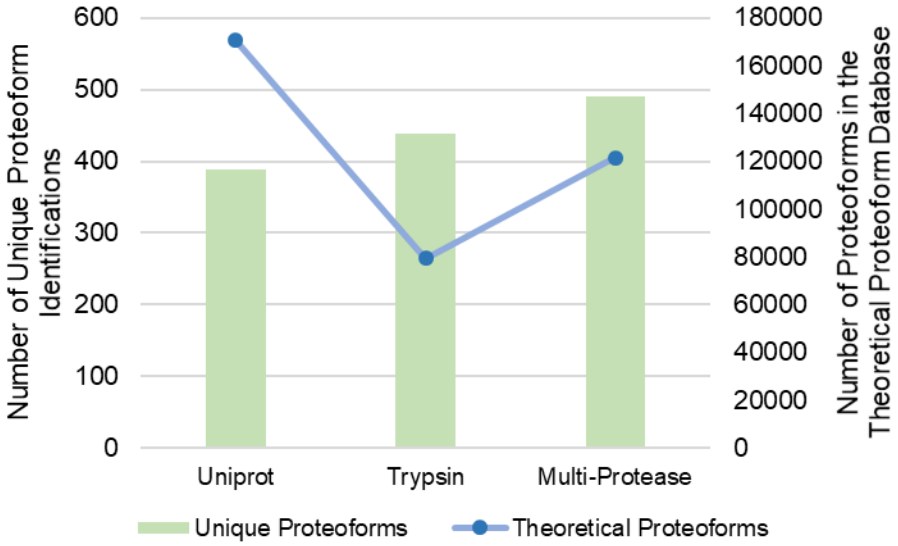


Figure 5: Comparison of proteoform results from searching the UniProt, trypsin protein pruned, multi-protease protein pruned databases. The number of proteoforms identified increase with the use of a protein pruned database and with the incorporation of multi-protease bottom-up data.

Aim 3: Elucidate the membrane proteoforms of pluripotent stem cells.

Generation of proteogenomic, multi-protease sample-specific database for pluripotent stem cells. Small aliquots of cells will be used for RNA extraction and RNA-Seq by procedures well developed in our lab.[71–74] A proteogenomic database including single amino acid variants (SAVs), novel splice junctions (NSJs), insertions and deletions (indels) will be constructed from the obtained RNA-Seq data using Spritz.

Multi-protease analysis of the cell membrane proteoforms with Arg-C, chymotrypsin, Glu-C, Lys-C and trypsin will generate a protein pruned database. Five aliquots of $1x10^8$ live cells will be oxidized with sodium periodate, tagged with biotin-SS-hydrazide, and lysed. Tagged proteoforms will be captured utilizing a streptavidin column and washed with high salt, high pH solutions to remove protein contaminants.[56,75–78] Captured proteoforms will be eluted from the column following reduction of biotin-SS-hydrazide's disulfide bond. Each aliquot will be divided into two sub-aliquots; one of which will be treated with PNGaseF while the other remains untreated, maintaining its intact glycosylations. Digestion of each sub-aliquot will occur using the filter-aided sample preparation (FASP) protocol[79] adjusted for each protease's optimal digestion conditions[72]. Nano-capillary low-pH LC-ESI-MS/MS and dynamic pH junction based CZE-ESI-MS/MS analysis will be performed using QE-HF orbitrap mass spectrometer, with a full MS scan being followed by ten MS/MS scans of the most intense ions. The box car mentioned in Aim 2 will be utilized to enhance sensitivity.

Multi-protease spectra will be analyzed with MetaMorpheus and all files will be searched against the proteogenomic sample-specific database. All files will be calibrated, novel PTMs will be discovered, peptides will be identified, and superior multi-protease protein inference will be performed to generate a comprehensive sample-specific multi-protease protein pruned database for improved proteoform analysis.

Top-down and intact mass analysis of pluripotent stem cell membrane proteins. Approximately $3x10^8$ pluripotent stem cells will be oxidized, tagged, and lysed to prepare for cell surface enrichment. Capture of the biotinylated proteoforms will occur utilizing a streptavidin conjugated column. Captured proteoforms will be washed with high salt, high pH solutions to eliminate non-membrane protein contaminants. [56,75–78] Proteoforms will be eluted following the reduction of the tag's disulfide bond with DTT. Proteoforms will then be fractionated by molecular weight using gel-eluted liquid fraction entrapment electrophoresis (GELFrEE). The approximately 150 ug of protein will be loaded onto a 10% GELFrEE cartridge and separated by molecular weight into 12 fractions according to manufacturer's instructions. Methanol-Chloroform extraction will be performed on each sample to remove incompatible detergents from the GELFrEE separation. Membrane proteins will be solubilized in optimized solvent conditions as determined in Aim 1. Samples will be analyzed by top-down HILIC-MS/MS and CZE-MS/MS, as well as intact mass HILIC-MS and CZE-MS. The boxcar strategy adapted for intact mass analysis and the top-down real-time precursor identification strategy will be utilized to increase proteoform identifications by increasing sensitivity of mass spectrometric analysis.

Top-down spectra will be analyzed with TDPortal utilizing the sample-specific database generated from multiple data streams. Intact mass files will be deconvoluted utilizing Thermo-Decon 4.0 and analyzed utilizing Proteoform Suite. The incorporation of the top down hit report generated from TDPortal and the sample specific database into Proteoform Suite's analysis of intact mass data will maximize the number of proteoform identifications possible.[80] Proteoform Suite will also be utilized to generate proteoform families.

Summary

Membrane proteins are underrepresented in "whole proteome" datasets due to their low abundance and insolubility. Solubilization conditions, enrichment techniques and the overall sensitivity of mass spectrometric analysis must be improved in order to facilitate membrane proteoform analysis. We propose the adaptation of conventional cell surface enrichment for proteoform analysis as well as the synthesis of a novel, cleavable, biotinylated hydrazide tag. We also propose the evaluation of membrane protein solubilization conditions and separation techniques to optimize delivery of proteoforms to the mass spectrometer. Advances in instrument control as well as the incorporation of multiple data streams for enhanced proteoform analysis will lead to increased sensitivity for proteoform analysis. All proposed improvements to the membrane proteoform analysis pipeline will be employed to catalogue the membrane proteoform-ome of pluripotent stem cells. This pipeline, in the future, could be utilized to profile membrane proteoform-ome of multiple cell and tissue types to generate a human membrane proteoform atlas. A study of this magnitude would increase the knowledge base surrounding membrane proteins as well as highlight the functional importance of various membrane proteoforms.

References:

1. Speers AE, Wu CC. Proteomics of Integral Membrane ProteinsTheory and Application. *Chem Rev*. 2007;107(8):3687-3714. doi:10.1021/cr068286z.

2. Wallin E, Heijne G Von. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*. 2010;7(4):1029-1038. doi:10.1002/pro.5560070420.

3. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol*. 2001;305(3):567-580. doi:10.1006/jmbi.2000.4315.

4. Ott CM, Lingappa VR. Integral membrane protein biosynthesis: why topology is hard to predict. *J Cell Sci*. 2002;115(Pt 10):2003-2009. http://www.ncbi.nlm.nih.gov/pubmed/11973342.

5. Torres J, Stevens TJ, Samsó M. Membrane proteins: The "Wild West" of structural biology. *Trends Biochem Sci*. 2003;28(3):137-144. doi:10.1016/S0968-0004(03)00026-4.

6. Zhou C, Zheng Y, Zhou Y. Structure prediction of membrane proteins. *Genomics, proteomics Bioinforma / Beijing Genomics Inst*. 2004;2(1):1-5. doi:10.1016/S1672-0229(04)02001-7.

7. Barrera NP, Robinson C V. Advances in the Mass Spectrometry of Membrane Proteins: From Individual Proteins to Intact Complexes. *Annu Rev Biochem*. 2011;80(1):247-271. doi:10.1146/annurev-biochem-062309-093307.

8. Ziegler YS, Moresco JJ, Tu PG, Yates JR, Nardulli AM. Proteomic analysis identifies highly expressed plasma membrane proteins for detection and therapeutic targeting of specific breast cancer subtypes. *Clin Proteomics*. 2018;15(1):1-9. doi:10.1186/s12014-018-9206-0.

9. Lu B, McClatchy DB, Jin YK, Yates JR. Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry. *Proteomics*. 2008;8(19):3947-3955. doi:10.1002/pmic.200800120.

10. Wu CC, MacCoss MJ, Howell KE, Yates JR. A method for the comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol*. 2003;21(5):532-538. doi:10.1038/nbt819.

11. Bausch-Fluck D, Hofmann A, Bock T, et al. A mass spectrometric-derived cell surface protein atlas. *PLoS One*. 2015;10(4):1-22. doi:10.1371/journal.pone.0121314.

12. Wollscheid B, Bausch-Fluck D, Henderson C, et al. Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol*. 2009;27(4):378-386. doi:10.1038/nbt.1532.

13. Smith LM, Kelleher NL, Linial M, et al. Proteoform: a single term describing protein complexity. *Nat Methods*. 2013;10(3):186-187. doi:10.1038/nmeth.2369.

14. Smith LM, Kelleher NL. Proteoforms as the next proteomics currency. *Science (80-)*. 2018;359(6380):1106-1107. doi:10.1126/science.aat1884.

15. Gregorich ZR, Ge Y. Top-down proteomics in health and disease: Challenges and opportunities. *Proteomics*. 2014;14(10):1195-1210. doi:10.1002/pmic.201300432.

16. Savaryn JP, Catherman AD, Thomas PM, Abecassis MM, Kelleher NL. The emergence of top-down proteomics in clinical research. *Genome Med*. 2013;5(6). doi:10.1186/gm457.

17. Siuti N, Kelleher NL. Decoding protein modifications using top-down mass spectrometry. *Nat Methods*. 2007;4(10):817-821. doi:10.1038/nmeth1097.

18. Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem*. 2016;9(1):499-519. doi:10.1146/annurev-anchem-071015-041550.

19. Chen B, Brown KA, Lin Z, Ge Y. Top-Down Proteomics: Ready for Prime Time? *Anal Chem*. 2018;90(1):110-127. doi:10.1021/acs.analchem.7b04747.

20. Cai W, Tucholski TM, Gregorich ZR, Ge Y. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev Proteomics*. 2016;13(8):717-730. doi:10.1080/14789450.2016.1209414.

21. Armirotti A, Damonte G. Achievements and perspectives of top-down proteomics. *Proteomics*. 2010;10(20):3566-3576. doi:10.1002/pmic.201000245.

22. Souda P, Ryan CM, Cramer WA, Whitelegge J. Profiling of integral membrane proteins and their post translational modifications using high-resolution mass spectrometry. *Methods*. 2011;55(4):330-336. doi:10.1016/j.ymeth.2011.09.019.

23. Abramson J, Ryan CM, Cramer WA, et al. Post-translational Modifications of Integral Membrane Proteins Resolved by Top-down Fourier Transform Mass Spectrometry with Collisionally Activated Dissociation. *Mol Cell Proteomics*. 2010;9(5):791-803. doi:10.1074/mcp.m900516-mcp200.

24. Zabrouskov V, Whitelegge JP. Increased coverage in the transmembrane domain with activated-ion electron capture dissociation for top-down fourier-transform mass spectrometry of integral membrane proteins. *J Proteome Res*. 2007;6(6):2205-2210. doi:10.1021/pr0607031.

25. Ahlf DR, Durbin KR, Catherman AD, et al. Large-scale Top-down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence. *Mol Cell Proteomics*. 2013;12(12):3465-3473. doi:10.1074/mcp.m113.030114.

26. Laganowsky A, Reading E, Hopper JTS, Robinson C V. Mass spectrometry of intact membrane protein complexes. *Nat Protoc*. 2013;8(4):639-651. doi:10.1038/nprot.2013.024.

27. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov*. 2002. doi:10.1038/nrd892.

28. La-Beck NM, Jean GW, Huynh C, Alzghari SK, Lowe DB. Immune Checkpoint Inhibitors: New Insights and Current Place in Cancer Therapy. *Pharmacotherapy*. 2015;35(10):963-976. doi:10.1002/phar.1643.

29. Link J. Direct analysis of large protein complexes using mass spectrometry. *Nat Biotechnol*. 1999;17:676-682. doi:10.1038/10890.

30. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data. *Mol Cell Proteomics*. 2005;4(10):1419-1440. doi:10.1074/mcp.R500012-MCP200.

31. Compton PD, Zamdborg L, Thomas PM, Kelleher NL. On the scalability and requirements of whole protein mass spectrometry. *Anal Chem*. 2011;83(17):6868-6874. doi:10.1021/ac2010795.

32. Weekes MP, Antrobus R, Lill JR, Duncan LM, Hör S, Lehner PJ. Comparative analysis of techniques to purify plasma membrane proteins. *J Biomol Tech*. 2010;21(3):108-115.

33. Smolders K, Lombaert N, Valkenborg D, Baggerman G, Arckens L. An effective plasma membrane proteomics approach for small tissue samples. *Sci Rep*. 2015;5:1-18. doi:10.1038/srep10917.

34. Takahashi N, Yamauchi Y, Izumi T, et al. Cell Surface Labeling and Mass Spectrometry Reveal Diversity of Cell Surface Markers and Signaling Molecules Expressed in Undifferentiated Mouse Embryonic Stem Cells. *Mol Cell Proteomics*. 2005;4(12):1968-1976. doi:10.1074/mcp.m500216-mcp200.

35. Wollscheid B, Mueller LN, Schiess R, Aebersold R, Schmidt A, Mueller M. Analysis of Cell Surface Proteome Changes via Label-free, Quantitative Mass Spectrometry. *Mol Cell Proteomics*. 2008;8(4):624-638. doi:10.1074/mcp.m800172-mcp200.

36. McDonald L, Robertson DHL, Hurst JL, Beynon RJ. Positional proteomics: Selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods*. 2005;2(12):955-957. doi:10.1038/nmeth811.

37. Shimkus M, Levy J, Herman T. A chemically cleavable biotinylated nucleotide: usefulness in the recovery of protein-DNA complexes from avidin affinity columns. *Proc Natl Acad Sci*. 2006. doi:10.1073/pnas.82.9.2593.

38. Zhao Y, Zhang W, Kho Y, Zhao Y. Proteomic Analysis of Integral Plasma Membrane Proteins. *Anal Chem*. 2004;76(7):1817-1823. doi:10.1021/ac0354037.

39. Zhang H, Li X jun, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol*. 2003. doi:10.1038/nbt827.

40. El-Faham A, Albericio F. Peptide Coupling Reagents, More than a Letter Soup. *Chem Rev*. 2011;111(11):6557-6602. doi:10.1021/cr100048w.

41. Hunter R, Caira M, Stellenboom N. Inexpensive, one-pot synthesis of unsymmetrical disulfides using 1-chlorobenzotriazole. *J Org Chem*. 2006;71(21):8268-8271. doi:10.1021/jo060693n.

42. Wilchek M, Bayer EA. Labeling Glycoconjugates with Hydrazide Reagents. *Methods Enzymol*. 1987;138(C):429-442. doi:10.1016/0076-6879(87)38037-1.

43. Lasserre JP, Beyne E, Pyndiah S, Lapaillerie D, Claverol S, Bonneu M. A complexomic study of Escherichia coli using two-dimensional blue native/SDS polyacrylamide gel electrophoresis. *Electrophoresis*. 2006;27(16):3306-3321. doi:10.1002/elps.200500912.

44. Graham RLJ, Graham C, McMullan G. Microbial proteomics: A mass spectrometry primer for biologists. *Microb Cell Fact*. 2007;6:1-14. doi:10.1186/1475-2859-6-26.

45. Goo YA, Ng W V., Hood L, et al. Proteomic Analysis of an Extreme Halophilic Archaeon, Halobacterium sp. NRC-1 . *Mol Cell Proteomics*. 2003;2(8):506-524. doi:10.1074/mcp.m300044-mcp200.

46. Klein C, Garcia-Rizo C, Bisle B, et al. The membrane proteome of Halobacterium salinarum. *Proteomics*. 2005;5(1):180-197. doi:10.1002/pmic.200400943.

47. Rahbar AM, Fenselau C. Integration of Jacobson's pellicle method into proteomic strategies for

plasma membrane proteins. *J Proteome Res*. 2004;3(6):1267-1277. doi:10.1021/pr040004t.

48. Yu YQ, Gilar M, Gebler JC. A complete peptide mapping of membrane proteins: A novel surfactant aiding the enzymatic digestion of bacteriorhodopsin [1]. *Rapid Commun Mass Spectrom*. 2004;18(6):711-715. doi:10.1002/rcm.1374.

49. Keller A, Melchior K, Tholey A, et al. Proteomic Study of Human Glioblastoma Multiforme Tissue Employing Complementary Two-Dimensional Liquid Chromatography- and Mass Spectrometry-Based Approaches. *J Proteome Res*. 2009;8(10):4604-4614. doi:10.1021/pr900420b.

50. Zhang H, Lin Q, Ponnusamy S, et al. Differential recovery of membrane proteins after extraction by aqueous methanol and trifluoroethanol. *Proteomics*. 2007;7(10):1654-1663. doi:10.1002/pmic.200600579.

51. Deshusses JMP, Burgess JA, Scherl A, et al. Exploitation of specific properties of trifluoroethanol for extraction and separation of membrane proteins. *Proteomics*. 2003;3(8):1418-1424. doi:10.1002/pmic.200300492.

52. Vattulainen I, Lee BW, Patra M, et al. Under the Influence of Alcohol: The Effect of Ethanol and Methanol on Lipid Bilayers. *Biophys J*. 2005;90(4):1121-1135. doi:10.1529/biophysj.105.062364.

53. Pinisetty D, Moldovan D, Devireddy R. The effect of methanol on lipid bilayers: An atomistic investigation. *Ann Biomed Eng*. 2006;34(9):1442-1451. doi:10.1007/s10439-006-9148-y.

54. Washburn MP, Wolters D. Nbt0301_242.Pdf. 2001;19(March). doi:10.1038/85686.

55. Liu H, Perkins PD, Boyes BE, Martosella J, Zolotarjova N, Moyer SC. High Recovery HPLC Separation of Lipid Rafts for Membrane Proteome Analysis. *J Proteome Res*. 2006;5(6):1301-1312. doi:10.1021/pr060051g.

56. Xenarios I, Langridge J, Vilbois F, Da Cruz S, Martinou J-C, Parone PA. Proteomic Analysis of the Mouse Liver Mitochondrial Inner Membrane. *J Biol Chem*. 2003;278(42):41566-41571. doi:10.1074/jbc.m304940200.

57. Vuckovic D, Dagley LF, Purcell AW, Emili A. Membrane proteomics by high performance liquid chromatography-tandem mass spectrometry: Analytical approaches and challenges. *Proteomics*. 2013;13(3-4):404-423. doi:10.1002/pmic.201200340.

58. Carroll J, Fearnley IM, Walker JE. Definition of the mitochondrial proteome by measurement of molecular masses of membrane proteins. *Proc Natl Acad Sci*. 2006;103(44):16170-16175. doi:10.1073/pnas.0607719103.

59. Gargano AFG, Roca LS, Fellers RT, Bocxe M, Domínguez-Vega E, Somsen GW. Capillary HILIC-MS: A New Tool for Sensitive Top-Down Proteomics. *Anal Chem*. 2018;90(11):6601-6609. doi:10.1021/acs.analchem.8b00382.

60. Walker SH, Carlisle BC, Muddiman DC. Systematic comparison of reverse phase and hydrophilic interaction liquid chromatography platforms for the analysis of N-linked glycans. *Anal Chem*. 2012;84(19):8198-8206. doi:10.1021/ac3012494.

61. Sun L, McCool EN, Kou Q, et al. Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome . *Anal Chem*. 2018;90(9):5529-5533.

doi:10.1021/acs.analchem.8b00693.

62. Cox J, Mann M, Geyer PE, Meier F, Virreira Winter S. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*. 2018;15(6):440-448. doi:10.1038/s41592-018-0003-5.

63. Wichmann C, Meier F, Winter SV, Brunner A, Cox J, Mann M. MaxQuant.Live enables global targeting of more than 25,000 peptides. *bioRxiv*. 2018:443838. doi:10.1101/443838.

64. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. *Nat Methods*. 2014;11(11):1114-1125. doi:10.1038/NMETH.3144.

65. Solntsev SK, Shortreed MR, Frey BL, Smith LM. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J Proteome Res*. 2018;17(5):1844-1851. doi:10.1021/acs.jproteome.7b00873.

66. Shortreed MR, Wenger CD, Frey BL, et al. Global identification of protein post-translational modifications in a single-pass database search. *J Proteome Res*. 2015;14(11):4714-4720. doi:10.1021/acs.jproteome.5b00599.

67. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res*. 2010;9(3):1323-1329. doi:10.1021/pr900863u.

68. Vit O, Petrak J. Integral membrane proteins in proteomics. How to break open the black box? *J Proteomics*. 2017;153:8-20. doi:10.1016/j.jprot.2016.08.006.

69. Miller RM, Millikin RJ, Hoffmann C V., et al. Improved Protein Inference From Multiple Protease Bottom-Up Mass Spectrometry Data. *Prep*.

70. Dai Y, Shortreed MR, Scalf M, et al. Elucidating Escherichia coli Proteoform Families Using Intact-Mass Proteomics and a Global PTM Discovery Database. *J Proteome Res*. 2017;16(11):4156-4165. doi:10.1021/acs.jproteome.7b00516.

71. Cesnik AJ, Shortreed MR, Sheynkman GM, Frey BL, Smith LM. Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. *J Proteome Res*. 2016;15(3):800-808. doi:10.1021/acs.jproteome.5b00817.

72. Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res*. 2014. doi:10.1021/pr4009207.

73. Sheynkman GM, Johnson JE, Jagtap PD, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics*. 2014;15(1). doi:10.1186/1471-2164-15-703.

74. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. *Mol Cell Proteomics*. 2013;12(8):2341-2353. doi:10.1074/mcp.O113.028142.

75. Schluesener D, Fischer F, Kruip J, Rögner M, Poetsch A. Mapping the membrane proteome of Corynebacterium glutamicum. *Proteomics*. 2005;5(5):1317-1330. doi:10.1002/pmic.200400993.

76. Zhang LJ, Wang XE, Peng X, et al. Proteomic analysis of low-abundant integral plasma membrane proteins based on gels. *Cell Mol Life Sci*. 2006;63(15):1790-1804. doi:10.1007/s00018-006-6126-

3.

77. Sickmann A, Berger C, Lewandrowski U, Moebius J, Zahedi RP, Walter U. The Human Platelet Membrane Proteome Reveals Several New Potential Membrane Proteins. *Mol Cell Proteomics*. 2005;4(11):1754-1761. doi:10.1074/mcp.m500209-mcp200.

78. Gatzeva-topalova PZ, Warner LR, Pardi A, Carlos M. A shotgun proteomic method for the identification of membrane-embedded proteins and peptides. *J Proteome Res*. 2011;18(11):1492-1501. doi:10.1016/j.str.2010.08.012.Structure.

79. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nat Methods*. 2009;6(5):359-362. doi:10.1038/nmeth.1322.

80. Schaffer L V., Shortreed MR, Cesnik AJ, et al. Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. *Anal Chem*. 2018;90(2):1325-1333. doi:10.1021/acs.analchem.7b04221.