METHODS DEVELOPMENT FOR GLYCOPEPTIDE AND GLYCAN ANALYSIS

By

Wijeweera Patabandige Milani Rasangika

Submitted to the graduate degree program in Chemistry and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Dr. Heather Desaire

Dr. David Weis

Dr. Steven Soper

Dr. Misha Barybin

Dr. Scott Hefty

Date Defended: September 7th, 2020

The dissertation committee for Wijeweera Patabandige Milani Rasangika
certifies that this is the approved version of the following dissertation:

METHODS DEVELOPMENT FOR GLYCOPEPTIDE AND GLYCAN ANALYSIS

Chairperson: Dr. Heather Desaire

Date Approved: September 24th, 2020

Abstract

Glycans introduce complexity to the proteins to which they are attached. These modifications vary during the progression of many diseases; thus, they serve as potential biomarkers for disease diagnosis and prognosis. The immense structural diversity of glycans makes glycosylation analysis and quantitation difficult. Therefore, a better understanding of various glycosylation profiling strategies; their strengths and weaknesses, is important towards selecting the best approach for a given clinical glycomics study. Not only that, successful application of glycomics analysis methods in the clinical glycomics field depends on using effective sample preparation strategies and better classification systems to accurately classify glycomics samples.

Among many analytical methods in the glycoproteomics analysis field, LC-MS analysis of glycopeptides is a frequent choice, as it provides information of both the glycans and their attachment sites. Numerous software tools have been developed to assist the glycopeptide identification workflow; however, those tools typically do a sub-optimal job when the glycopeptides of interest are in low abundance or when they are poorly ionized. Therefore, in such incidences, expert targeted analysis approaches, where LC-MS data is manually interpreted to confidently identify the recalcitrant glycopeptides would be beneficial. Thus, chapter 2 of this dissertation introduces a simple, expert analysis method, the peak alignment approach, which relies on high-resolution MS data and chromatographic retention times to assign the glycosylation sites. The method identifies a set of co-eluting glycopeptides in an LC-MS experiment using a reverse phase column; these glycopeptides are extracted based on a limited *N*-linked glycan library, and once the co-eluting glycopeptides are identified, they are verified by using high-resolution MS data and confirmed by using MS/MS data. The developed method successfully quantified many of the glycosylation sites of a heavily glycosylated human plasma glycoprotein within a single LC-MS run while requiring less sample amount and less analysis time, compared to the state-of-the-art competing analysis method.

Sample preparation is a vital step in all glycomics analysis studies, as it affects both the sensitivity and the selectivity of the analysis. Altered glycosylation of specific proteins can serve as a biomarker for diverse diseases. Uromodulin is one such glycoprotein; it is a biomarker for kidney health. Current strategies of uromodulin glycosylation analysis are time-consuming and tedious; they involve complex steps to enrich uromodulin, label glycans, followed by post sample clean-up, which limit the utility of these methods in clinical glycomics studies. Therefore, chapter 3 of this dissertation introduces a simple and straightforward direct ESI-MS analysis performed in the negative ion mode to quantify *N*-linked glycans of uromodulin, enriched from urine samples of two different biological states. The developed method enriches uromodulin directly from urine via ultrafiltration performed with a 50 kD molecular weight cut-off filter; it omits any labeling steps that require post-sample clean-up and includes steps to reduce the salt contents of the samples to minimize the ion suppression during the direct ESI-MS detection. The method proved to be highly reproducible over multiple samples' preparations and over multiple analyses; it was useful for accurately quantifying uromodulin glycans and classifying the samples of different biological states into clearly distinguishable groups by PCA.

Sample classification based on the whole glycomic profile, instead of selecting a single glycan feature or a few glycan features, could benefit the sample classification through identifying underlying trends of the glycomics data. The Aristotle Classifier is one such supervised classification algorithm that uses not only all the individual glycans abundances, but also their relative proportions to each other, to classify samples. Once this classifier was built, its’ classification ability needed to be challenged and compared with standard classification methods, like PCA. However, acquiring large sets of real glycomics samples with known glycosylation differences is difficult; thus, we chemically generated large sets of IgG glycomics data in-house, to mimic two different biological states as healthy and disease. Therefore, chapter 4 of this dissertation describes the optimization of both the sample preparation and LC-MS conditions to generate large sets of IgG glycopeptides’ data to mimic samples of a healthy state and a disease state. Of these samples, the healthy state was represented by samples with a native IgG glycosylation profile while the disease state was represented by samples with a slightly altered IgG glycosylation profile. The generated data were quantified, but the samples could not be classified into healthy versus disease based on any individual glycopeptide of the samples. Therefore, the data proved to be challenging; thus, they were submitted to both the Aristotle Classifier and to a principal components analysis (PCA), to challenge each approach’s classification ability. The generated results showed that the Aristotle Classifier outperformed the PCA classification in multiple data sets.

Acknowledgments

Family is the most precious gift that anyone could ever imagine. I am blessed with a nice family where I found the courage to fight through many obstacles during my journey towards perusing a PhD. Therefore, my first thank goes to my mother, who always stayed by my side like a shadow when the life was too harsh. In addition, I would like to thank my father, brother, sister, my nephew and niece for listening to me and filling my days with happiness. Also, I want to thank my husband, a great partner and a true friend, who was so patient with me throughout the past five years, enduring a long distance relationship, until I fulfil my dreams.

Thank you to the faculty of Chemistry, the University of Kansas, for teaching me necessary scientific knowledge that significantly benefitted towards making progress in my research projects. My utmost thanks goes to my supervisor, Dr. Heather Desaire, for being a great mentor. She has provided me her kind guidance and support to overcome numerous difficulties I have faced during the past few years as a graduate student. In addition, she allowed me to work as an intern at the Food and Drug Administration (FDA), where I explored the field of released glycan analysis of biotherapeutics. I am so lucky to have you as my mentor and I will forever thankful for your endless kindness.

Furthermore, I would like to thank previous and current Desaire group lab members who supported me. My special thank goes to Dr. Eden P. Go, Dr. Kasun Imaduwage, Dr. Josh Shipmann, Dr. Wenting Hue, and Dr. Jude Lakbub for helping me with the instrument training and subject related matters. Also, for Hanna and Leah for keeping me company.

Finally, I would like to thank my second family, Lawrence Sri Lankan community, who made my life at KU easier, happier, and healthier. I will never forget the support they have offered me during this toughest time period of my life.

Table of Contents

Chapter 1 . Introduction

1.1 Protein Glycosylation

Protein glycosylation is the most complex post-translational modification, and more than 50% of human proteins are glycosylated. The process of protein glycosylation occurs within the endoplasmic reticulum and Golgi apparatus, and it is controlled by a series of enzymes that modify the carbohydrates that are covalently attached to proteins through certain amino acid residues.[1-3] This modification is complex to study, in part, because of the heterogeneous nature of the glycans. Unlike protein biosynthesis, glycan biosynthesis does not rely on an underlying template; thus, the resultant glycan structures can be very heterogeneous. Both the enzyme availability and the cellular environment can affect the final glycosylation profile.[2] In addition, this complexity is further enhanced by the presence of multiple monosaccharide units, which are linked together in a variety of ways to form glycan structures; glycans can have various compositions, and even differently-linked isomers with identical composition, due to variety in linkage and branching.[4-5]

These heterogeneous glycans (oligosaccharides) attached on proteins play crucial roles in regulating various biological processes such as fertilization,[6-7] protein folding and stabilization,[8-9] cellular recognition, cellular adhesion,[10-11] and immune defense.[12] In addition, glycosylation is considered as a critical quality attribute in biotherapeutics production, since the glycans attached on proteins greatly affect the safety and the efficacy of protein-based drugs. Thus, a minor change in glycosylation profile of these drugs can lead to serious conditions, such as adverse immune reactions,[13-14] rapid clearance,[15-16] and loss of therapeutic potency. Furthermore, aberrant glycosylation of various endogenous proteins have been associated with the progression of diseases, such as cancers,[17-19] kidney diseases[20] among others; thus, glycans may serve as clinical biomarkers for disease diagnosis and prognosis.[21] Therefore, deeper understanding of this complex modification, protein *N*-linked glycosylation, and current *N*-linked glycosylation profiling strategies, is critical, not only to identify sensitive biomarkers, but also to provide information necessary to regulate the glycosylation in biotherapeutics, so the safety and the activity of glycoprotein-based drugs is ensured.

1.1.1 *N*-linked Glycosylation

The most common glycosylation type, *N*-linked glycosylation, occurs when the glycans are attached to the proteins through the amide nitrogen on the side chain of an Asn residue. These glycosylated Asn are usually located within a unique amino acid sequence: Asn-Xxx-Ser/Thr, where Xxx can be any amino acid except proline.[1-3, 22] $Glc_3Man_9GlcNAc_2$ is the common building block for all the *N*-linked glycans, and this precursor is attached to the protein during the initial phase of the glycosylation process, as shown in Figure 1. This precursor undergoes many enzymatic trimming and monosaccharide addition steps that introduce modifications to the precursor glycan while preserving the tri-mannosyl-pentasaccharide core ($Man_3GlcNAc_2$).[2-3] These modifications to the precursor glycan result three major types of *N*-linked glycan structures; they are high-mannose (Man), complex, and hybrid. (See Figure 1.)

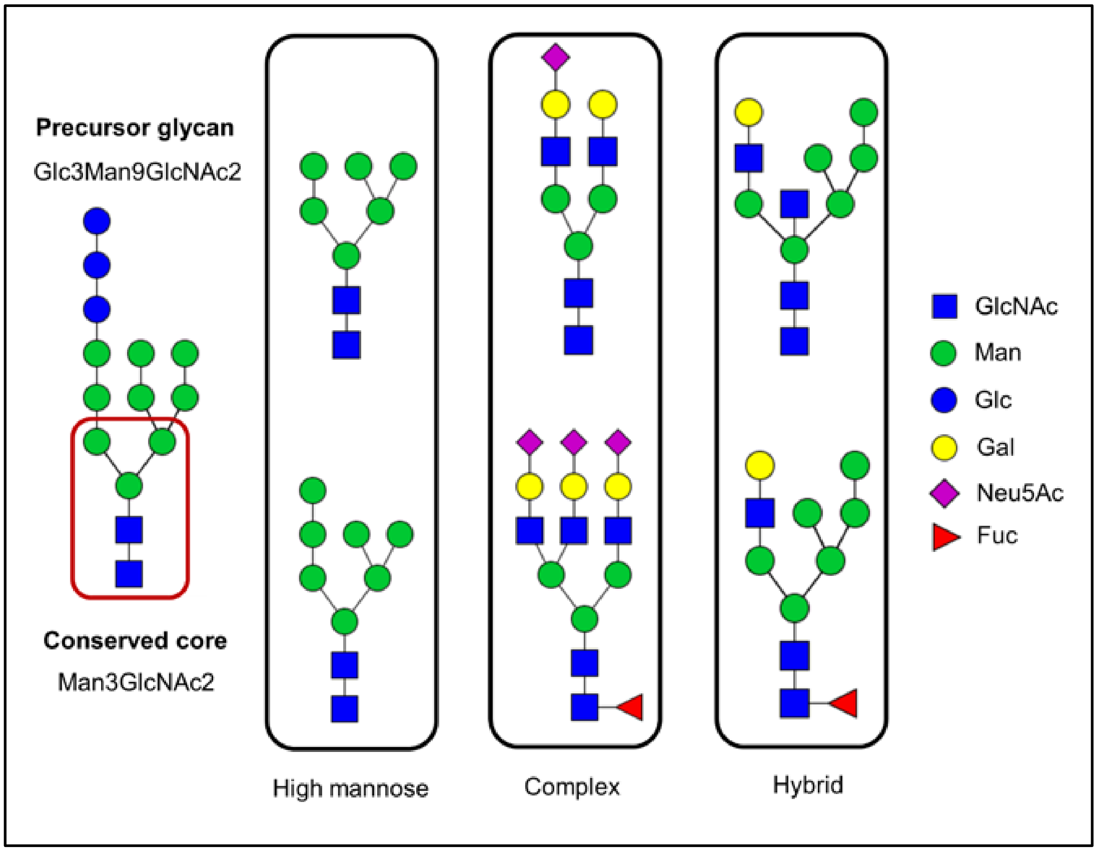


Figure 1. Symbolic representation of different *N*-linked glycans. High mannose, complex and hybrid are three major *N*-linked glycan types; all derived from the common precursor Glc3Man9GlcNAc2. These glycans are made of *N*-acetyl glucosamine (GlcNAc), mannose (Man), glucose (Glc), Galactose (Gal), *N*-acetyl neuraminic acid (Neu5Ac), and fucose (Fuc).

The high-mannose type glycans are formed by trimming of monosaccharides from the precursor without addition of new monosaccharides, thus leaving only Man residues attached to the core structure. In contrast, complex- type glycans are formed by trimming monosaccharides of the precursor glycan, followed by addition of new sugars to the terminal mannose residues of both arms of the $Man_3GlcNAc_2$ core. In complex type structures, GlcNAc is the very first monosaccharide unit directly linked to the terminal mannoses in the core-structure, and it is further extended with additional monosaccharides; the most common pattern involves attachment of galactose (Gal) units and terminal sialic acid (*N*-acetylneuraminic acid) units. Based on the number of GlcNAc attached to the terminal mannose sugars in the core structure of the complex-type glycans, the number of branches are defined as bi-, tri- and tetra-antennary. In addition, in complex-type glycans, core-fucosylation and/or antennary fucosylation can be observed when the fucose (Fuc) is attached to the innermost GlcNAc of the core structure or the GlcNAc at the non-reducing end. The hybrid-type glycans, the last glycan category, share the characteristic features of both high-mannose and complex-type glycan structures.[2-3]

1.1.2 Altered *N*-linked Glycosylation and Diseases

During the progression of many diseases, alteration of the *N*-linked glycosylation profile is observed. These alterations may include both upregulation and/or downregulation of glycans, elevated branching, size increase, and modifications to the core-structure.[20, 23-26] For instance, differentially expressed serum IgG *N*-linked glycans, with decreased levels of high-mannose structures, reduced levels of core-fucosylation and sialylation were observed during colorectal cancer progression,[23] while decreased levels of fully galactosylated *N*-linked glycan structures were identified in gastric cancers,[24] lung adenocarcinoma tissues,[25] and rheumatoid arthritis (RA).[27] In addition, significantly decreased levels of Man5 and bi-antennary *N*-linked glycans, along with elevated levels of branching, antennary fucosylation, and core-fucosylation were observed in the serum glycans of primary epithelial ovarian cancer patients.[26] The altered glycans and glycosylation patterns that are unique to certain types of diseases may serve as biomarkers, and discovery of those biomarkers is important, not only to understand disease pathology, but also to perform more selective treatments and disease diagnoses.

In the past few decades, impressive efforts have been made to identify clinically relevant glycan biomarkers for diseases. More than 90 potential *N*-linked glycan biomarkers have been identified based on previously published studies, including biomarkers for certain types of cancers, such as breast, liver, ovarian, kidney, and pancreatic cancers, as well as for Hepatitis B and C, Alzheimer's disease, and diabetes.[28] This large number of potential glycan-based biomarkers clearly show the significance of quantitative glycomic studies in discovering selective candidate glycan biomarkers for distinguishing disease states from healthy states, and also in disease prognosis, diagnosis and/or treatment. However, the discovery of unique biomarkers for various diseases is greatly dependent on not only the availability of sensitive and reliable analytical methods, but also on careful selection of the most appropriate and cost-effective approach for any clinical glycomics study. Thus, in this chapter, we compare the performance of four commonly used quantitative glycomics methods to guide the selection of an appropriate analytical strategy for a given clinical and pre-clinical study.

1.2 General Considerations

1.2.1 Glycosylation Analysis

Glycosylation analysis can be performed in two ways: glycopeptide-based analysis and glycan-based analysis.[2-3, 29] During glycopeptide-based analysis, the glycans remain attached to the glycosylation site, and therefore, retain information about the protein to which they are attached and the site of attachment. This information increases the specificity of the analysis, but the trade-off is that the analyses are more complex. Site-specific glycosylation analysis is used limitedly in the glycan biomarker discovery field,[30] due to the glycopeptides' lower abundance, lower ionization efficiency, need of method optimization for each glycoprotein,[31] and difficulties in data interpretation.[32] By contrast, glycan-based analysis, where glycans are released from the glycoproteins and then are analyzed, is useful for obtaining aggregate information about the total glycan pool. While the method provides substantially less specificity, it is frequently employed in clinical glycomics due to the availability of universal and well-established protocols for glycan analysis. One way to balance the strengths of both methods is to perform glycan analysis on a specific protein target that has been enriched from the sample. In this case, the glycan analysis provides information specific to the protein of interest, and well-established methods can be used to facilitate quantitation and analysis.

1.2.2 Sample Preparation

Figure 2. shows multiple ways of generating glycans or glycopeptides from complex biological samples for glycosylation analysis.

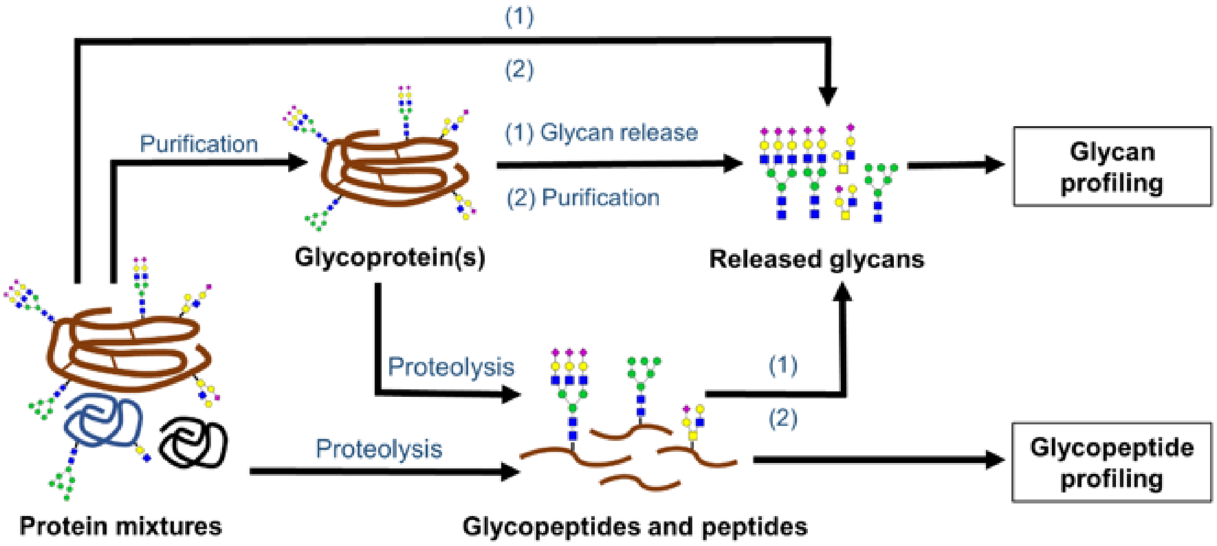


Figure 2. Different sample preparation routes for generating glycans and glycopeptides from complex biological samples for quantitative glycosylation analysis. In this workflow, for glycan profiling; glycans are released from purified glycoprotein(s), or directly from the crude biological mixture. Alternatively, glycans can be released from glycopeptides. For glycopeptide profiling; glycopeptides that are generated from either the purified glycoprotein(s) or directly from complex biological mixtures (for high abundant glycoproteins) are subjected to proteolysis; then, the resulting mixture of glycopeptides and peptides are subjected to quantitative analysis.

Of these sample preparation steps, glycoprotein purification at the crude protein mixture level is performed especially when a targeted quantitation of a specific protein's glycome profile is necessary; for example, IgG is affinity purified with protein A or G, prior to the quantitation of IgG *N*-linked glycans associated with cancer in serum samples.[23-24] On the other hand, glycans can be released directly from the non-enriched biological samples when the total glycome pool of a biological matrix is quantified;[33] however, the method yields limited specificity. In glycopeptide analysis, proteolysis at crude mixture level is performed when the targeted glycoprotein is in high abundance;[30, 34] but, the proteolysis on purified glycoprotein(s)[32, 35-36] is used more frequently as it improves both the sensitivity and the specificity of the analysis. Once these glycans or glycopeptides are generated, further purification can be performed by solid phase extraction-based methods (SPE) with porous graphitized carbon (PGC) [33, 37] and hydrophilic interaction liquid chromatography (HILIC),[32, 38-39] or by using specific lectins for glycopeptides.[40]

1.2.3 Quantitation

Glycan abundances from healthy patients versus those of a disease state are compared by either absolute[4, 37] or relative quantitation;[4, 17, 37] relative quantitation being the more common choice, since absolute quantitation usually requires glycan standards that are not readily available.[41] In relative quantitation, the proportion of glycans present in the two sample types (healthy versus disease state) is reported by dividing an individual glycan abundance by the total glycan abundance,[42-43] or by the abundance of the most intense glycan peak,[31, 44] or by a peak among the major signals.[31] While these methods do not report exact glycan concentrations, the ratios measured typically are useful for allowing the identification of under- or over-expressed glycans between healthy versus disease groups, which is the ultimate goal of the analysis.

1.3 Quantitative Strategies in Clinical Glycomics

1.3.1 Mass Spectrometry (MS)-Based Approaches

MS-based approaches are widely used in clinical and pre-clinical glycomic studies. This choice is preferred by many researchers because the method is sensitive, and it can be used to differentiate species with unique masses. Structural information can also be obtained through MS/MS and MS^n experiments.[45-49] These benefits, especially when coupled with separation and enrichment techniques, facilitate the identification and quantitation of glycans originating from complex biological matrices. Matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) and electrospray ionization mass spectrometry (ESI-MS) are the most common ionization techniques for glycan analysis, and the latter may be done in conjunction with widely used separation techniques, such as liquid chromatography and capillary electrophoresis, while the former requires offline separation.

1.3.1.1 MALDI-TOF MS Analysis of Released *N*-linked Glycans

MALDI-TOF MS is widely applied in quantitative clinical glycomics. It is a simple, sensitive, high-throughput method.[20, 23, 33, 37, 50-51] The most common sample preparation procedure for glycan analysis by MALDI-TOF MS is the workflow appearing at the top of Figure 2: the desired glycoprotein of interest is isolated and purified from the complex biological matrix; next, release of *N*-linked glycans from glycoproteins via PNGase F treatment typically follows. The glycans are subsequently purified and labeled [20, 23, 33, 37, 51] for MALDI-TOF MS analysis. Figure 3A shows more detail about the MALDI-MS specific sample preparation steps.

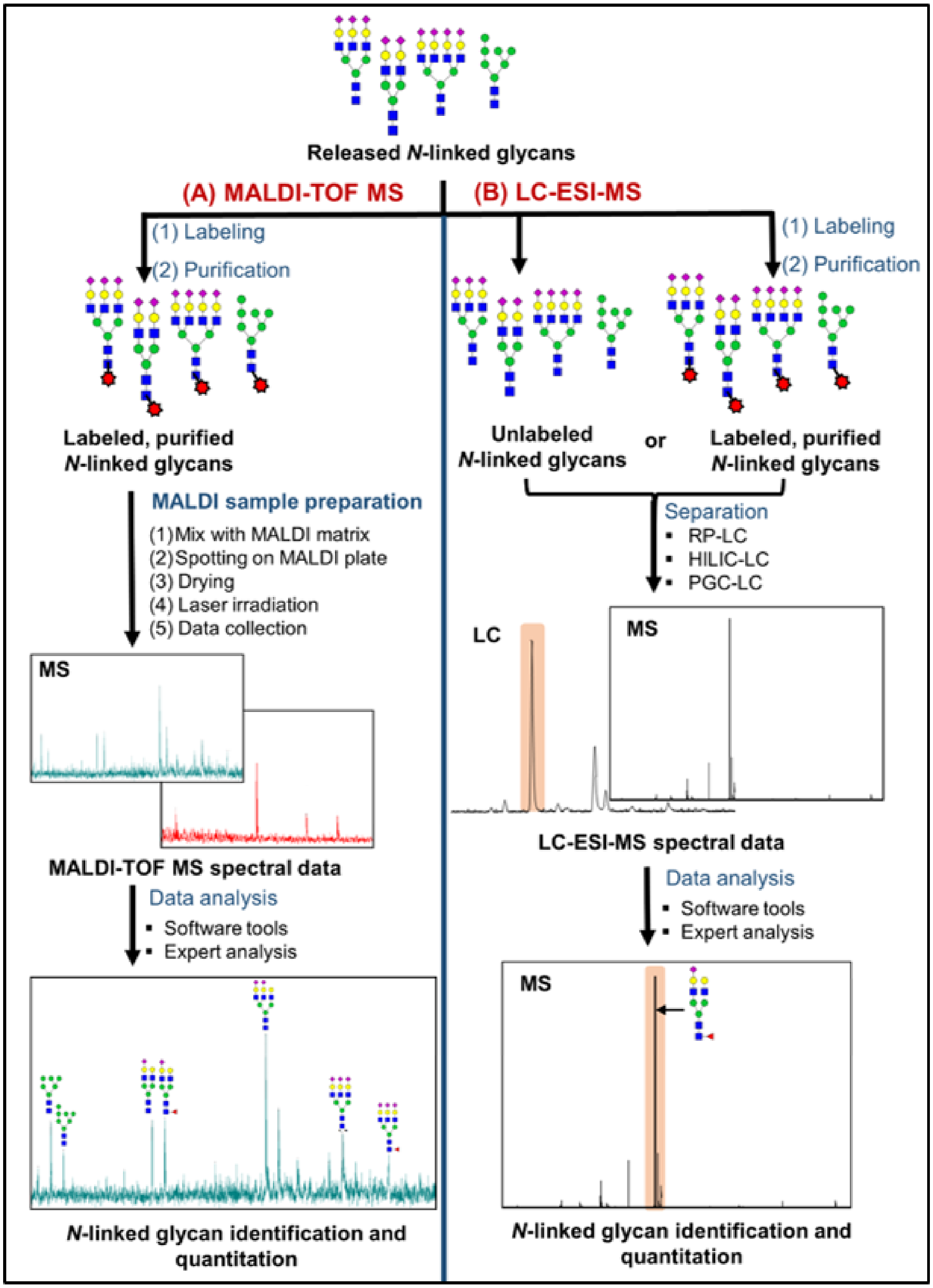


Figure 3. General workflows for Released *N*-linked glycan quantitation by MALDI-TOF MS **(A)** and by LC-ESI-MS **(B)**. In MALDI-TOF MS analysis, labeled *N*-linked glycans are mixed with a MALDI-matrix and irradiated with laser shots to collect MS data **(A)**. In LC-ESI-MS, labeled or unlabeled *N*-linked glycans are separated by using a liquid chromatography method, followed by ESI-MS analysis **(B)**.

Once the glycans are purified, labeling of released *N*-linked glycans is usually performed to enhance the ionization efficiency of glycans, to improve the sensitivity of the analysis, and sometimes to allow simultaneous detection of both neutral and acidic *N*-linked glycans.[23, 51] MALDI-TOF MS analysis of permethylated *N*-linked glycans[20, 33, 37, 50] has been performed in many clinical glycomics studies. In addition, derivatization of sialic acid via methyl esterification[43] or ethyl esterification[38] is another useful labelling method. Once the labeled glycans are purified again; they are prepared for MALDI-TOF MS analysis.

Sample-matrix preparation greatly affects the quality of the resultant MS spectra, as matrix plays a vital role in promoting solid phase analytes into the gaseous phase. Therefore, prior to MALDI-TOF MS analysis, the labeled glycans are mixed in a 1:1 ratio with a MALDI matrix, such as 2,5-dihydroxy-benzoic acid (2,5-DHB),[38, 52] Super DHB,[20] 4-chloro-α-cyanocinnamic acid (Cl-CCA),[52] 2,4,6-trihydroxyacetophenone (THAP),[52] or 9-Aminoacridine (9-AA),[53] followed by spotting the aliquots of the mixed sample solutions onto a MALDI plate. Then multiple laser shots are applied on each sample spot to ionize the samples, followed by the MS analysis.[20, 23, 33, 37, 51] During the analysis, reflectron-positive[20, 37-38, 54-55] and negative,[52] as well as the linear-positive and negative ion modes[52] are used. Among them, positive-ion mode is more commonly used due to higher ionization efficiency and higher S/N ratio reported for labeled glycans.[29] However, negative ion mode is also used to detect acidic glycans with improved detection sensitivity.[52]

1.3.1.2 LC-ESI-MS Analysis of Released *N*-linked Glycans

ESI-MS is also used in quantitative glycomics studies. A general LC-ESI-MS/MS workflow of glycan quantitation includes: glycan release, purification, labeling, followed by glycan separation, and ESI-MS quantitation. Similar to MALDI-TOF MS, in LC-ESI-MS-based glycan quantitation; glycan release is commonly performed at the crude mixture level for total glycome quantitation, followed by enrichment of released glycans using SPE.[4, 25] However, as mentioned above, performing glycan release after enriching for a target protein gives protein-specific information; this approach affords the opportunity of providing larger differences between disease states and healthy states when glycans from a disease-impacted protein are selected for profiling.

Figure 3B represents the general workflow for LC-ESI-MS analysis. Once the glycans are purified, analysis can be performed on labeled, [33, 48, 56-58] unlabeled,[4, 24] or chemically reduced[25, 59] glycans. Glycan labeling is performed to improve the sample throughput, by allowing for multiple, differentially labeled samples to be analyzed together, and to enhance the ionization. Isobaric tags or tandem mass tags (TMT)[48, 56-58] that have identical masses, but with various heavy isotopes distributed within the structure, are commonly used in labeling experiments. AminoxyTMT,[48, 58] is one such tag that allows simultaneous labeling of glycans derived from multiple samples; resulting in a single chromatographic peak at the full MS level for various glycans, yielding sample-specific reporter ions at the MS/MS level for comparative *N*-linked glycan quantitation. Alternatively, stable isotopic labeling of glycans where small mass differences to the glycans derived from multiple samples are introduced through isotopically labeled reducing-end labeling agents, or isotopic permethylation, are used to quantify glycans at the MS1 level.[56] Permethylation improves the sensitivity of the analysis by enhancing ionization, while allowing simultaneous detection of neutral and acidic glycans;[60] and when incorporated with isotopic labeling, it improves the throughput of the analysis. For example, 8-plex quantitative glycan analysis of multiple breast cancer cell lines was feasible through isotopic permethylation; performed by using isotopically labeled iodomethane during the permethylation reaction.[56] Furthermore, metabolic isotope labeling, where cells from different samples are labeled isotopically, is also used in quantitative glycomics analyses to reliably quantify glycans from different samples while minimizing potential sample preparation bias.[33] At the end of the appropriate glycan labeling step, the samples are purified from the labeling agent and/or biological matrix molecules, and they are analyzed by MS.

Glycan separation prior to the MS detection facilitates the enrichment of various glycan structures derived from complex biological samples, while allowing sensitive detection of multiple glycans by MS. Liquid chromatography (LC) is used frequently, due to its ability to resolve complex mixtures, its compatibility with MS methods, and its capacity for facilitating automation. As compared to the traditional LC-ESI-MS or MALDI-TOF MS, nano-LC-MS is used in many studies, as it significantly improves the detection sensitivity.[4] Nano-LC columns packed with C_{18}[56, 60-61] or PGC-bonded stationary phases[58] are frequently used to separate permethylated *N*-linked glycans, while HILIC columns are used to separate more hydrophilic glycans.[48] Alternatively, by incorporating a microfluidic chip to the nano-LC workflow, a greater retention time reproducibility, better separation, and high sensitivity for the glycans can be achieved.[4, 24] For example, PGC chip-nano-LC separation was used in many quantitative glycomic studies, as PGC is capable of separating glycans by their polarity, size, and the 3D structure, while exhibiting good isomer separation capacity.[4, 24-25, 59] Once the glycans are effectively separated, they are detected with MS for quantitation.

Electrospray ionization (ESI) is a commonly used ionization technique in quantitative glycomics studies because it generates glycan ions without the loss of labile groups; thus, it provides complete composition information. In many studies where ESI-MS is used, the instrument is operated in a data dependent mode,[57-58, 61] acquiring full MS scans in an Orbitrap,[57,

[61] for example, followed by MS/MS scans of the most intense ions. Also, in some studies, targeted quantitation is performed using a triple quadrupole mass spectrometer operated in multiple reaction monitoring (MRM) mode.[61] Once the LC-MS data are acquired, they are analyzed by using software tools or by combining both expert analysis and automated tools prior to the *N*-linked glycan quantitation.

1.3.1.3 LC-MS Analysis of Glycopeptides

LC-MS analysis of glycopeptides is another method used in biomarker discovery; this approach provides glycosylation site-specific information. However, the method is challenging because it involves determination of both an unknown peptide and an unknown glycan. The general workflow for glycopeptide analysis using LC-MS includes: isolation of desired glycoprotein(s) from the biological matrices, glycoprotein denaturation, reduction and alkylation, all prior to the enzymatic digestion; then separation of enriched or non-enriched glycopeptides is achieved, usually by HPLC, followed by MS analysis.

In this workflow, isolation of glycoprotein(s) at either the crude mixture level[32] or at glycopeptide level[35-36] is important due to the glycopeptides' low abundance as compared to the non-glycosylated peptides. Of these enrichment methods; lectin-based enrichment, where carbohydrate-binding proteins or lectins bind to specific carbohydrate moieties via affinity binding, is widely implemented for both glycoprotein level[62] and glyopeptide level enrichment.[40, 46, 63] For example, *Aleuria aurantia* lectin (AAL) [62] and *Lens culinaris* Agglutinin (LCA)[46] are two lectins that bind to fucosylated carbohydrate motifs, and *Sambucus nigra* lectin (SNA) is another lectin that binds to the carbohydrates that contain sialic acids.[62] Furthermore, unique antibodies that identify specific epitopes present on protein backbones are also used to isolate glycoproteins from complex biological samples.[32, 64] Anti-IgA[64] and antihuman Haptoglobin

(Hp)[32] are two such antibodies that used to isolate IgA and haptoglobin, derived from cancer-associated serum samples. Additionally, protein G or protein A-based isolation of IgG[64-65] and glycopeptide level enrichment with sepharose beads[65] are also reported. However, when the targeted glycoprotein is in high abundance, such as immunoglobulin G in serum, enrichment steps at either the crude mixture level or glycopeptide level can be avoided, while performing proteolytic digestion directly on crude biological samples, as shown in Figure 2.[30, 34]

To generate glycopeptides, the (enriched) samples are subjected to proteolytic digestion with site-specific proteases, such as trypsin,[30, 66] or non-specific proteases, or a combination of both.[32, 34, 41] The most common means of generating glycopeptides is to use proteases that have high specificity, such as trypsin. Site-specific proteases are useful because they produce peptides that can be predicted in advance, and the number of different peptides generated per glycosylation site is very limited; often a single unique peptide will be generated per glycosylation site. Some researchers, however, are concerned that specific proteases limit the number of glycopeptides identifed when the resultant glycopeptides are multiply glycosylated or miscleaved. They support the use of non-specific proteases, which may be better in the specific cases of multiply glycosylated and difficult-to-digest peptides. This strategy, either on its own, or in combination with specific proteases, may be useful in obtaining more complete glycopeptide identification in some cases.[34] After protease digestion, the resulting glycopeptide mixture, which is subjected to purification/enrichment[35, 63, 66] or not,[30, 34] is typically analyzed by LC-MS, as shown in Figure 4.

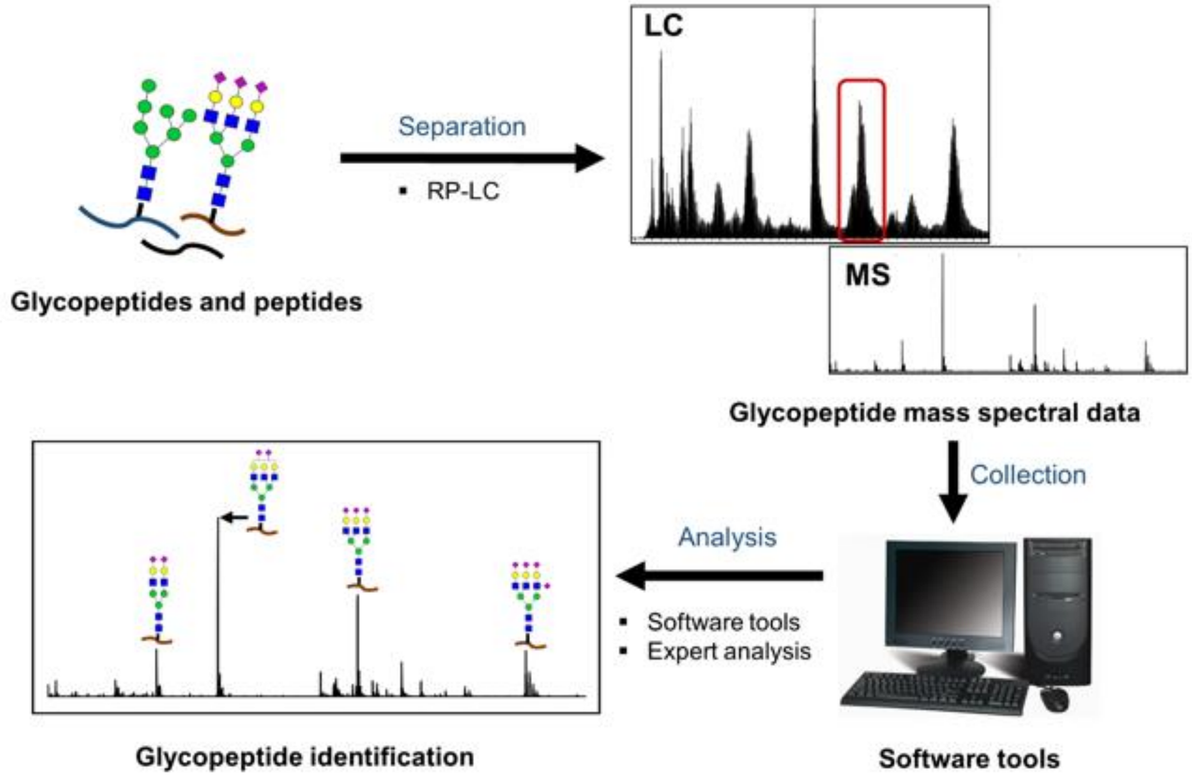


Figure 4. General workflow for LC-MS analysis of glycopeptides. In this workflow, enriched or non-enriched mixtures of glycopeptides generated at purified glycoproteins level or at crude mixture level are further separated followed by the mass spectrometry detection and analysis for glycopeptide identification and quantification.

Glycopeptide separation prior to MS detection is important because it permits enrichment of glycopeptides from peptide counterparts that co-exist in the mixture; those can reduce the ionization of glycopeptides if they co-elute. LC is the method of choice for glycopeptide separation owing to its MS compatibility, glycopeptide resolving capacity, and ability to be automated.[3] Reverse phase (RPLC) columns with C_{18}- bonded phases are the most popular in glycopeptide separation;[30, 36, 40-41, 66-72] these columns separate glycopeptides based on the interactions between the peptide backbone and the hydrophobic stationary phase rather than the glycan interactions.

Once the glycopeptides are separated, they are detected by using MS. Multiple reaction monitoring (MRM) mode, a targeted mass spectrometry approach, is frequently used for quantifying glycopeptides due to its high sensitivity and selectivity.[30, 34, 63, 73] Alternatively, untargeted approaches are also possible; in these cases, glycopeptides can be quantified by their high-resolution ESI-MS signal[74-75] or by using data-dependent LC-MS/MS.[62] ESI with positive ionization mode is frequently used, but in some cases, negative ion mode is also performed to enhance the ionization of sialylated glycopeptides.[40] Finally, the data analysis is performed by using either software tools[30, 34] or by combining both software tools and manual verification.[32]

1.3.2 Spectroscopy-Based Approaches

1.3.2.1 LC/CE-Fluorescence Profiling of Released *N*-linked Glycans

Another widely used method in quantitative clinical glycomics studies is HPLC with fluorescence detection. In this case, labeled glycans are analyzed. As with previously discussed methods, this one also requires isolation and purification of glycoprotein(s) of interest from complex biological samples, prior to sample preparation; see Figure 2. Then the isolated glycoprotein(s), which are either denatured[76] or non-denatured[20] are subjected to enzymatic glycan release with PNGase F enzyme, followed by glycan purification, which can be done using HILIC-SPE.[31] The samples undergo fluorescent labeling prior to analysis.[20, 77-78] Some common fluorescent labeling reagents include 2-aminobenzoic acid (2-AA)[38] and 2-aminobenzamide (2-AB).[20, 43, 77-78] Once the glycans are fluorescently labeled, the excess labeling reagent is removed using SPE,[31, 43-44, 76] paper chromatography,[77, 79] or size-exclusion chromatography[80] and the labeled glycans are separated by HPLC. Options include HILIC-LC (hydrophilic interaction liquid chromatography),[43-44] HPAEC (high pH anion exchange chromatography),[20, 78] and NP-HPLC (normal phase high performance liquid chromatography).[42] The separated, derivatized glycans are quantified based on their fluorescence signal. See Figure 5 for a representative workflow for this method.

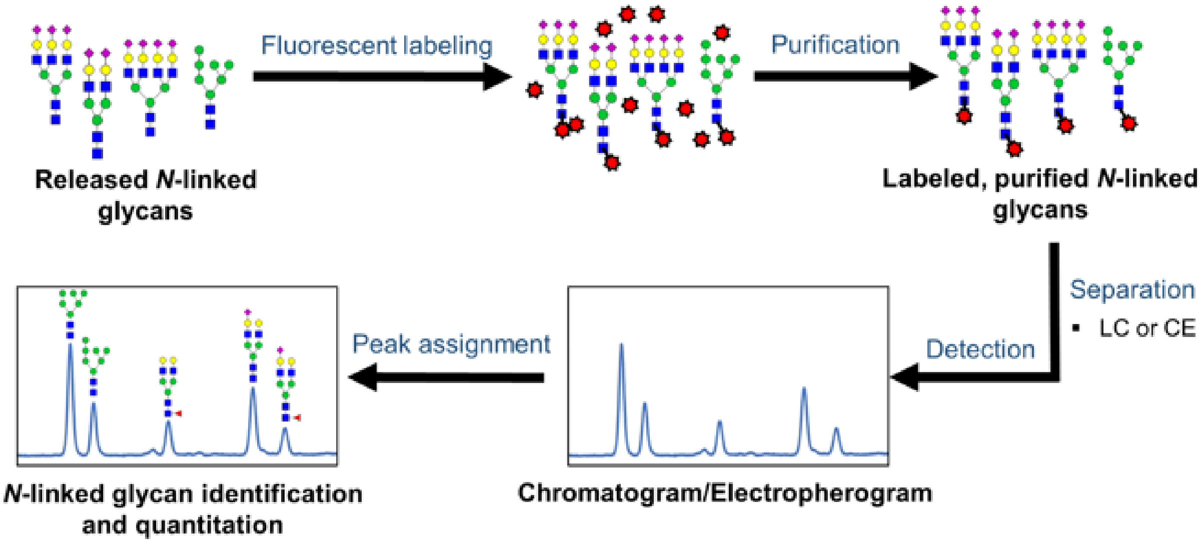

Released *N*-linked glycans
Fluorescent labeling
Purification
Labeled, purified *N*-linked glycans
Separation
▪ LC or CE
Detection
Chromatogram/Electropherogram
Peak assignment
N-linked glycan identification and quantitation

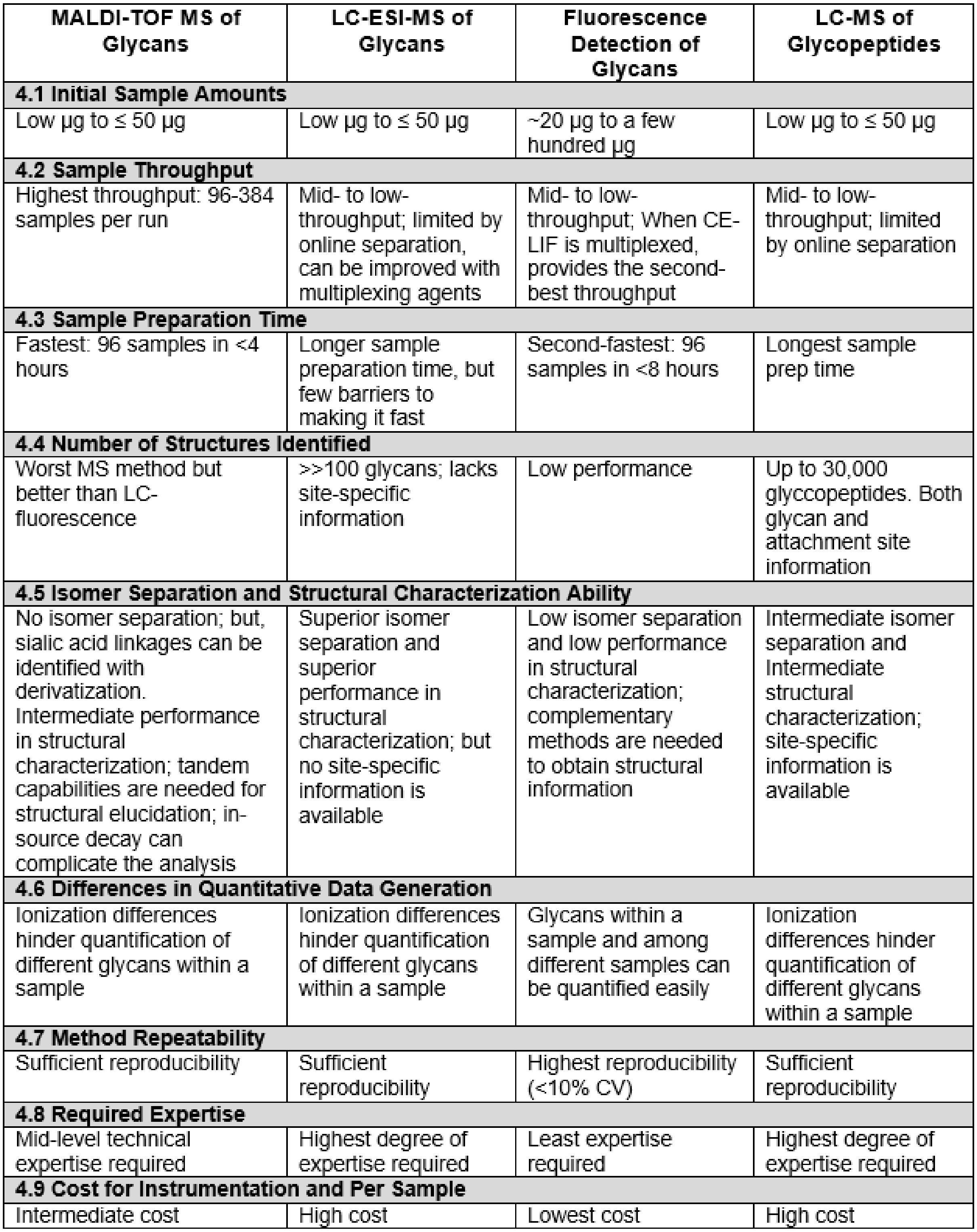
Figure 5. Workflow for the analysis of released *N*-linked glycans via liquid chromatography (LC)/ capillary electrophoresis (CE)-fluorescence detection. In this workflow, released *N*-linked glycans are labeled with a suitable fluorescence labeling reagent, purified, and then are separated by using LC or CE. Finally, the resultant peaks of chromatograms or electropherograms are assigned by using established data bases and/or follow-up experiments followed by *N*-linked glycans quantitation.

Capillary electrophoresis (CE) separation paired with fluorescence detection is also used to profile *N*-linked glycans in clinical glycomics studies, as the method is high-throughput and readily adaptable to microfluidic devices.[55, 80-81] The sample preparation for CE and HPLC are similar; however, 8-aminopyrene-1,3,6-trisulfonic acid (APTS), [20, 31, 43, 77, 82] and 8-aminonaphthalene-1,3,6-trisulfonic acid (ANTS)[83] labels are often used; these labels carry three negative charges from sulfonic acids. The charges on fluorophores increase the electrophoretic separation.[84] However, when the glycans carry negatively-charged sialic acid groups, before the APTS labeling, cleavage of the sialic acids[81] or neutralization via chemical modifications such as methylamidation[55] is performed. This modification yields glycans all bearing the same charge state, thus, allowing glycans' electrophoretic migrations to be based on their hydrodynamic volumes; resulting increased migration times for sialylated glycans while preserving the efficiency of separation.[55] Once the glycans are labeled, they are separated by using various CE modes, such as conventional CE[55, 80] or capillary gel electrophoresis (CGE);[43, 77] then are detected with fluorescence.

1.4 Performance Comparison

As described in the previous section, various glycomic analysis and quantitation strategies have been developed; each of these methods are currently used in glycomics-based biomarker studies. However, each of these methods has differences in their workflows and unique advantages and disadvantages. Thus, one must carefully consider when to use a particular quantitative strategy for a biomarker study. To assist in this selection process, herein we compared the performance of the four quantitative strategies described above with respect to various key figures of merit. Table 1 summarizes the methods' performance.

Table 1. Performance comparison of four glycomics analysis platforms.



MALDI-TOF MS of Glycans	LC-ESI-MS of Glycans	Fluorescence Detection of Glycans	LC-MS of Glycopeptides
4.1 Initial Sample Amounts			
Low µg to ≤ 50 µg	Low µg to ≤ 50 µg	~20 µg to a few hundred µg	Low µg to ≤ 50 µg
4.2 Sample Throughput			
Highest throughput: 96-384 samples per run	Mid- to low-throughput; limited by online separation, can be improved with multiplexing agents	Mid- to low-throughput; When CE-LIF is multiplexed, provides the second-best throughput	Mid- to low-throughput; limited by online separation
4.3 Sample Preparation Time			
Fastest: 96 samples in <4 hours	Longer sample preparation time, but few barriers to making it fast	Second-fastest: 96 samples in <8 hours	Longest sample prep time
4.4 Number of Structures Identified			
Worst MS method but better than LC-fluorescence	>>100 glycans; lacks site-specific information	Low performance	Up to 30,000 glyccopeptides. Both glycan and attachment site information
4.5 Isomer Separation and Structural Characterization Ability			
No isomer separation; but, sialic acid linkages can be identified with derivatization. Intermediate performance in structural characterization; tandem capabilities are needed for structural elucidation; in-source decay can complicate the analysis	Superior isomer separation and superior performance in structural characterization; but no site-specific information is available	Low isomer separation and low performance in structural characterization; complementary methods are needed to obtain structural information	Intermediate isomer separation and Intermediate structural characterization; site-specific information is available
4.6 Differences in Quantitative Data Generation			
Ionization differences hinder quantification of different glycans within a sample	Ionization differences hinder quantification of different glycans within a sample	Glycans within a sample and among different samples can be quantified easily	Ionization differences hinder quantification of different glycans within a sample
4.7 Method Repeatability			
Sufficient reproducibility	Sufficient reproducibility	Highest reproducibility (<10% CV)	Sufficient reproducibility
4.8 Required Expertise			
Mid-level technical expertise required	Highest degree of expertise required	Least expertise required	Highest degree of expertise required
4.9 Cost for Instrumentation and Per Sample			
Intermediate cost	High cost	Lowest cost	High cost

1.4.1 Sample Size

In complex biological mixtures, the glycoprotein(s) of interest are usually present at low abundance compared to the non-glycosylated proteins. Therefore, the detection of these low abundant glycans/glycopeptides depends on not only the method sensitivity, but also the initial sample amount used for the analysis. Table 1. shows the comparison of methods in terms of typically used initial glycoprotein amounts for the quantitative glycomics analysis.

Generally, MS-based methods are highly sensitive; these methods can be implemented with lower microgram quantities of glycoproteins to provide reliable quantitative glycomics data. When LC-MS-based analysis of glycopeptides/glycans is implemented, higher quantitation sensitivity is achieved by using different MS-based strategies, such as targeted multiple reaction monitoring (MRM)[30, 34, 60] or advanced tandem MS techniques;[32, 65] these methods permit lower initial sample amounts. For examples, two studies performed by Hong *et al*.[30, 34] used MRM-mode to quantify immunoglobulin glycopeptides generated from approximately 24 µg of IgG,[30, 34] 5 µg of IgA,[34] and 3 µg of IgM,[34] derived from 2 µL of un-enriched serum samples; the glycoprotein amounts indicated in here are approximate values calculated based on these glycoproteins' average plasma concentrations reported in the literature.[85] In addition, MS/MS techniques, such as LC-ESI-MS/MS[65] and LC-EThcD MS/MS[32] were used to quantify glycopeptides of IgG (~24 µg),[65] and haptoglobin (3 µg)[32] derived from serum samples of pancreatic cancer and liver cirrhosis patients, respectively. LC-MS analysis of released *N*-linked glycans is also performed at lower microgram levels;[24, 60] one study quantified more than 55 permethylated serum *N*-linked glycans, by using approximately 0.1 µg of total glycoproteins derived from 0.005 µL of enriched-serum sample injected on column;[60] yet readers should be aware that the amount of glycoprotein initially subjected to sample preparation for these studies was more than 100 times greater than the reported injection amounts. Similar to LC-MS analysis methods, MALDI-TOF MS analysis also uses initial glycoprotein amounts ranging from 0.5 – 30 µg, derived from up to 5 µL of biological samples for *N*-linked glycan quantitation.[20, 23, 37, 86] The study performed by Gao and coworkers[86] quantified more than 50 TMPP-Ac-labeled *N*-linked glycan structures derived from approximately 0.5 µg of serum glycoproteins per MALDI spot; however, the starting quantities, prior to sample preparation, were 20 times higher.

Compared to all MS-based methods, fluorescence-based methods typically require comparatively higher initial glycoprotein amounts, ranging from approximately twenty to a few hundred micrograms.[76, 80]

1.4.2 Sample Throughput

High-throughput methods that are capable of analyzing glycomic profiles of several thousands of biological samples are necessary to perform large-scale clinical studies.[42] The throughput of MALDI-TOF MS is superior to all other glyco-analytical techniques, and it is followed by CE-LIF and HPLC-FLD.[39, 43, 77, 82] See Table 1 for an abbreviated comparison.

To assess the throughput of profiling human plasma IgG *N*-glycosylation of 1201 individuals, four platforms were compared: MALDI-TOF MS, LC-ESI-MS, and two spectroscopic approaches.[77] MS analysis was performed on purified tryptic glycopeptides, while non-mass spectrometric methods, UPLC-HILIC-FLR and multiplex capillary gel electrophoresis-laser induced fluorescence detection (xCGE-LIF), were performed on 2-AB and APTS labeled, released *N*-linked glycans, respectively.[77] MALDI-TOF MS proved to be far superior, while LC-ESI-MS was the slowest. In another example, HILIC-UHPLC-FLD, xCGE-LIF and MALDI-TOF MS approaches were compared for identifying the serum *N*-glycome changes associated with the rheumatoid arthritis and pregnancy. Again, MALDI-TOF MS sample throughput outperformed the spectroscopic methods.[43, 77] These studies showed that 96-384 samples could be analyzed by MALDI within a single run, providing the measurement of a sample at a sub-minute time scale.[77, 87]

Apart from MALDI-TOF MS, CGE-LIF, when multiplexed, proves to be the method with the second-best throughput. It allows for the analysis of thousands of samples within a day.[43, 77] The typical run time for either the CGE-LIF or HPLC-FLD is approximately in 40 – 60 min range, but once CGE is multiplexed with up to 96 capillaries in parallel, the required analysis time per sample can be reduced to the low minute scale.[43] As compared with CGE-LIF, the throughput of conventional CE-LIF is lower as it lacks multiplexing ability, but the typical run time is generally lower than both CGE-LIF and UPLC-FLR.

UPLC-FLR and LC-ESI-MS show medium throughput;[77] the throughput of both of these methods are limited by the front-end gradient time. For example, one study quantified total plasma *N*-linked glycan profiles obtained from 2396 individuals by using an HPLC-FLD. The reported total analysis time was 106 min per sample.[44]

While LC-ESI-MS/MS analysis of glycans is one of the slowest methods, sample throughput can be improved by using different multiplexing agents, such as tandem mass tags (TMT) or isobaric labels,[58] and stable isotopic labels.[56] Introducing multiplexing agents is useful not only for improving the reliability of the quantitation, but also for increasing the number of samples that can be analyzed within a single LC-MS run, while lowering the analysis time per sample.[58] Recently, sixplex AminoxyTMT mass tags were used by Merchef and co-workers[58] to reliably quantify serum *N*-linked glycans derived from individuals with various esophageal diseases. They quantified 44 glycans after labeling them with TMT sixplex reagents, followed by glycan separation on a PGC column and analysis with nano-LC-ESI-MS/MS.[58] In addition, a very recent study performed by Li and coworkers[88] presented mass-defect-based, duplex-dimethyl pyrimidinyl ornithine (DiPyrO) tags with a mass difference of 45.3 mDa at the MS1 level; these tags were used to quantify *N*-glycome profile differences of human serum samples derived from cancer patients before and after chemotherapy; the study permitted quantification of 36 glycans, presented at relatively high abundance in the control samples as compared to the samples collected after chemotherapy.[88]

1.4.3 Sample Preparation Time

Sample preparation is required for all these methods because of the complexity of the sample matrix and the heterogeneity within the sample. Typically, on a glycoprotein, a variety of glycans can be attached to either a single glycosylation site or to multiple glycosylation sites found on the peptide backbone. This heterogeneity results many different protein glycoforms, which are usually in low abundance compared to the non-glycosylated proteins present in the biological matrix.[2-3] Therefore, efficient glycoprotein purification and separation at the crude-mixture level or the glycan/glycopeptide level is vital in glycomics analysis, as any contaminant present in the sample can affect the detection sensitivity, reproducibility, and relative glycan quantitation.[76] Therefore, many improvements in glycoprotein purification, sample preparation, including release of *N*-linked glycans, glycan enrichment, and labeling, have been reported in the literature; these methods are described in sections 1.3.1.1, 1.3.1.2, 1.3.1.3 and 1.3.2.1. However, the complexity of these multiple glycan/glycopeptide processing steps directly affects the total sample preparation time of the analysis; making it difficult to perform large-scale clinical studies on disease-related glycan biomarkers.[80]

The throughput of preparing samples for HPLC-FLD, CE-LIF and MALDI-TOF MS analysis of released *N*-linked glycans is quite similar.[43] All of these methods include glycoprotein enrichment, enzymatic *N*-linked glycan release performed overnight, glycan derivatization, and detection of purified glycans. However, the sample preparation throughput of MALDI-TOF MS currently surpasses the non-MS based methods. This is well-evidenced by two previous studies performed by Shubhakar *et al*.[50] and Bladergroen *et al*.;[89] they have presented high-throughput, clinically-feasible, automated sample preparation for MALDI-TOF MS analysis; these automated protocols allowed 96 clinical samples to be processed and detected within about 7 h for permethylated samples [50] and 3.5 h for samples with sialic acids esterified.[89] During these studies, the sample preparation workflow was expedited through introducing robotic liquid handling systems, which significantly reduced the sample preparation time for the analysis. On the other hand, automation of non-MS-based sample preparation, for example HPLC-FLD, has been also reported; one of these studies reduced the sample processing time for 96 2-AB labeled samples from 72 h[79] to 22 h[76] by introducing an alternative approach to the conventional in-gel block method; the new method supported full automation of sample preparation for the glycosylation analysis of an individual glycoprotein. Another study done by the same group[90] reported processing of 100 samples within approximately 14 hours, by using the same method, but with some modifications that permitted whole serum glycan analysis.

Compared to MALDI-TOF MS and fluorescence-based methods, relatively little effort has been directed at speeding up sample preparation for LC-ESI-MS of released glycans and LC-MS of glycopeptides, and these methods typically have lower sample preparation throughput.[91] In principle, sample preparation for LC-ESI-MS analysis of glycans would be approximately equivalent to that of MALDI-TOF MS, since the same types of analytes are studied. Yet efforts to demonstrate expedient sample preparation methods for ESI-MS based analysis have not been published, likely because the slow step becomes the instrument throughput, not the sample prep.

In contrast to analyzing released glycans, LC-MS analysis of glycopeptides requires different sample preparation steps that must be done in advance: glycoprotein enrichment, denaturation, reduction, alkylation, and enzymatic digestion. The sample preparation time consumes about 3 hours prior to the enzymatic digestion; enzymatic digestion is typically performed overnight.[32] The time allotted to the LC-MS/MS analysis can vary from 20 min[34, 73] to a few hours,[41] which reduces the analytical throughput of the method.

1.4.4 Number of Structures Identified

Identification of as many as unique glycan/glycopeptide species present in biological samples is important in biomarker research; the more structures quantified, the greater the likelihood that researchers will be able to identify glycans whose abundance changes with the disease sate.

Among different MS-based methods, LC-MS analysis of glycopeptides is capable of identifying the highest number of analytes per analysis. For example, one recent study quantified more than 600 glycopeptides across over 50 serum glycoproteins by implementing a dynamic multiple-reaction monitoring (dMRM) method optimized in a UHPLC-QqQ; the study permitted quantitation of sialylated, fucosylated glycans, in addition to low abundant high mannose-type glycans.[41] In another study, Kazuhiro *et al*.[40] identified more than 30 000 AAL-affinity-enriched glycopeptides derived from serum samples of hepatocellular carcinoma (HCC) patients, chronic hepatitis patients, and healthy controls via LC-TOF-MS while allowing the identification of multifucosylated glycans of alpha-1-acid glycoprotein (AGP), as a candidate HCC biomarker.

LC-ESI-MS analysis of released *N*-linked glycans, provides the second-best coverage of glycosylated analytes. Among many studies where a higher number of glycan identifications were reported, PGC-chip-based separation was used. Song *et al*.[59] performed an analysis on reduced serum *N*-linked glycans and identified more than 170 *N*-linked glycan structures, including complete (50) and partially elucidated (100) structures that were included in a representative library for serum. Moreover, in another study, out of 115 glycan structures identified, 29 were altered in lung adenocarcinoma tissue samples as compared to the non-malignant tissues.[25]

Among all MS-based methods, MALDI-TOF MS shows the lowest coverage of unique glycans. However, compared to fluorescence-based methods, MALDI-TOF MS is capable of assigning more glycan compositions, and it provides good separation for more complex tri- and tetra-antennary glycan structures.[43]

In fluorescence-based methods, the number of unique species detected is limited[43, 55, 92] and the assignment of each individual analyte peak requires prior knowledge about the retention time or the migration time of the species being analyzed. Table 1 summarizes the different methods' capacities.

1.4.5 Isomer Separation and Structural Characterization Ability

Glycomics analysis is complicated because of structural diversity introduced by different glycan compositions, linkages, and branching patterns.[4-5] Accurate identification of numerous glycan compositions and in-depth structural characterization of different glycans or glycopeptides structures, including isomers, is very important due to their biological significance, diagnostic relevance, and biotherapeutic importance.[5] As indicated in Table 1, LC-MS-based methods are thus far the most informative for structural assignment of glycans/glycopeptides; however, combination of both the optimized separation strategies and tandem MS techniques, are required to perform isomer separation and structural characterization.[5]

Considerable research has been invested into achieving isomeric separation for released glycans, followed by characterization by tandem mass spectrometry. Porous graphitized carbon (PGC),[4, 93-94] hydrophilic-interaction liquid chromatography (HILIC),[95] and reversed-phase (RP)-LC[61] are potential choices for the stationary phases for isomer separation, while tandem MS techniques, such as collision induced dissociation (CID)[59, 61, 93-94] and higher energy collision dissociation (HCD) [97-98] are main choices for glycans' structural characterization. Two recent studies performed by Yehia Mechref and coworkers,[93-94] used a PGC-nLC-MS/MS method performed at higher temperature (75 °C), to effectively separate and characterize permethylated glycans derived from multiple cancer cell lines. This study allowed efficient separation of glycan structures including many different glycan isomers; such as monosaccharide positional isomers (core- or branched fucose and α3- or α6-branched galactose) and linkage isomers; these isomers were then effectively identified by using specific diagnostic fragment ions observed in tandem MS spectra. Overall, these studies allowed identification of more than 100 glycan isomer structures derived from less than 50 glycan compositions. Apart from the frequently used PGC-nLC-MS/MS, RP-nLC-MS/MS is also used for glycan isomer separation and characterization; in one example, permethylated *N*-linked glycans from HCC patients were characterized, and 82 potential isomeric glycans from 52 glycan compositions were identified.[61] However, the use of RP-LC for *N*-linked glycan isomer separation is limited due to the poor resolution observed for permethylated isomeric glycans, and the poor retention observed for hydrophilic glycans.[93] In addition to RP-LC, HILIC is also used to separate *N*-linked glycan isomers, for an example, linkage isomers of ProA-labeled sialylated glycans.[95]

Isomer separation at the glycopeptide level is also important, as it permits the quantitation of site-specific isomeric glycan alterations. Generally, RP-LC is the method of choice for glycopeptide separation; it successfully separates multiple glycopetides with different peptide backbones; but, it poorly resolves the glycan isomers those all have a common peptide backbone.[96-97] Therefore, RP-LC has been used limitedly in glycopeptides' isomer-specific studies; one such example is the study performed by Yuan *et al.*;[47] they used RP-nLC-MRM-MS to quantify linkage-specific fucosylation differences of *N*-linked glycopeptides of seven plasma-derived glycoproteins of liver cirrhosis patients; these authors used outer arm fucosylation-specific fragment ion(s) in the tandem MS spectra for targeted transitions; they found that increased outer arm fucosylation is related to the progression of disease. Instead of reverse phase chromatography, HILIC separation obtained significant attention in recent years because of their glycopeptide isomer separation ability.[96-97] Two recent studies performed by Huang *et al*.[96] and van der Burgt *et al*.[97] used HILIC-LC in combined with targeted MS approaches to separate glycopeptide isomers of human IgG[96] and prostate-specific antigen (PSA);[97] both of these studies allowed differentiation of linkage-specific sialylated glycopeptides, and also resolved galactose position of G1F glycan of IgG glycopeptides.[96] Though, glycopeptide-based analysis provides site-specific information, it typically shows poor isomer separation ability as compared to the LC-ESI-MS analysis of released *N*-linked glycans while providing limited glycan-specific structural information. Therefore, if the goal of the study is to obtain comprehensive structural information of various glycans; the best choice would be the LC-ESI-MS of released glycans, which enables effective separation and in-detail characterization of glycan structures including many different isomers.

As compared to LC-MS-based methods discussed in this chapter, MALDI-TOF MS lacks the ability to separate glycan isomers as the method is not supported by front end glycan separation; thus, typically provide glycans' compositional assignments, but not the isomer-specific information.[43] However, MALDI-TOF MS by itself permits sialic acid linkage differentiation, only when sialic acids are subjected to linkage-specific derivatization prior to the analysis.[38, 43, 98] For an example, Reiding *et al*.[38] identified 77 plasma *N*-linked glycan compositions belonging to 108 glycan structures, after subjecting sialic acid α-2,6 and α-2,3 linkages to ethyl esterification and lactonization, respectively. In another study, MALDI-TOF MS was used to identify differences in sialylation linkages of ethyl esterified serum *N*-linked glycans derived from the samples of normal pregnant individuals and those with rheumatoid arthritis.[43] Moreover, though MALDI-TOF MS is useful for assigning different glycan compositions, structural elucidation of those compositions requires additional tandem capabilities, and also, the structural assignment of glycans can be complicated due to the loss of labile groups as a result of in-source decay.[99]

Many studies have compared the glycan isomer separation capacity of MALDI-TOF MS and non-MS-based approaches. Among various methods discussed in this chapter, CE-LIF and UPLC-FLR methods also allow for effective separation of *N*-linked glycans while permitting branch-specific information and separation of various isomers.[43, 77, 82] HILIC-FLR and CGE-LIF methods are able to distinguish between the 3-arm and 6-arm galactosylation differences, in addition to the fucose position (core- or branched-) of fucosylated glycans.[43, 77] By contrast, during these studies, MALDI-TOF MS analysis was not able to provide isomer-specific information for these glycans.

However, fluorescence is not a method that is well-suited to provide structural information about the glycans in a sample, as it primarily is used for quantitation. When using fluorescence to quantify glycans, other methods or tools need to be paired with it if information about the glycan is needed. For example, well-characterized glycan standards can be used to match the retention times in LC-fluorescence analyses; or additional follow-up MS experiments[44, 55, 84] could be done; or sequential enzymatic digestion can be used [39, 42-43, 84] to obtain structural information. These additional methods, which need to be done in conjunction with fluorescence-based quantification, introduce limitations when one of the researchers' goals is to identify the structure(s) of the glycan(s) that interest them.

1.4.6 Differences in Quantitative Data Generation

MS-based approaches used in glycomic quantitation are more complex than LC/ CE-fluorescence-based methods. In MS-based methods, the glycan response (peak abundances) are affected by both structural composition of the glycans and as well as the co-eluting interferences that suppress the ionization of glycans or glycopeptides (LC-MS based approaches). Therefore, the relative abundances of the glycans or glycopeptides cannot be compared across different compositions within the same sample. However, in LC/CE-fluorescence methods, glycan labeling is stoichiometric and is not affected by the nature of the glycan type or composition. In these methods, fluorescence dye is attached only to the reducing end of the glycan, and none of the structural differences of *N*-linked glycans are found at this end.[42] Therefore, it is possible to assume that all the labeled *N*-linked glycans fluoresce with a similar quantum yield, while allowing reliable quantitation of glycan peak areas in the same sample and between samples.[42, 77] Because of this point, LC/CE-fluorescence based methods are preferable if the research study requires that the relative quantities of glycans within a sample to be measured accurately.

1.4.7 Method Repeatability

Method repeatability is an important factor that needs to be considered during quantitative clinical glycomics studies where many sample sets are being analyzed. When the method is highly reproducible; it permits improved sensitivity, thus, allowing for differentiation of minor changes occurring in multiple samples.[34]

Many studies show that the repeatability of LC-FLD analysis of released *N*-linked glycans is superior to all other analytical methods.[43-44, 76] Typically, HPLC-FLD yields lower than 10% coefficient of variation, especially for major glycan peaks of the sample,[43-44, 76] and even a lower CV value (1.6%) was reported for the 10 most abundant *N*-linked glycans derived from plasma samples of RA patients.[43] CE-LIF analysis of glycans is also highly reproducible, but it is second to the LC-FLD method.[31, 43]

In MALDI-TOF MS, the reproducibility is affected by not only the variation of the analyte ionization but also the spot-to-spot variation of the laser pulse. Therefore, compared to the non-MS based methods, MALDI-TOF MS analysis reported the least reproducible glycan quantitation data in many studies;[43, 89] these studies showed that the reproducibility of the analysis can be improved when quantifying glycan-derived traits instead of individual glycan peaks.

Similar to MALDI-TOF MS, other MS-based methods also show lower repeatability as a result of both LC run-to-run variation and the ionization differences that occur during the MS analysis.[58] Throughout the literature, the use of MS-based methods with sufficient repeatability (CV<15%) for the quantitative clinical glycomics analysis have been reported for both LC-MS analysis of glycopeptides and released *N*-linked glycans. Lebrilla and colleagues[34, 41] and Shih *et al.*[65] reported a less than 15% intra-day and inter-day repeatability for serum glycopeptide quantitation. Similarly, for LC-ESI-MS *N*-linked glycan quantitation, sufficient reproducibility was reported in many studies with lower run-to-run variation and over multiple sample preparations.[58-59]

1.4.8 Required Expertise

MS-based methods typically require higher expertise compared to non-MS based methods; both the operation of the mass spectrometer and the more complex data analysis require significant experience.[43, 77] Among the MS-based methods, MALDI-TOF MS is the most straightforward, but for LC-ESI-MS analysis of both glycopeptides and released *N*-linked glycans, the required expertise is very high; researchers not only have to have a solid command of mass spectrometry, but also HPLC separation. Additionally, ESI data is often more complicated to analyze, especially if it is from glycopeptides. In contrast to MS-based methods, the primary skill necessary to perform UPLC-FLR and CE-LIF methods is expertise in separations. While these methods also require training for sample preparation and instrument handling, well-established glycan preparation protocols are available; a straight forward detection method and well-established data bases also simplify data analysis for fluorescence-based methods. [43, 76-77]

1.4.9 Cost for Instrumentation and per Sample

Typically, the instrumentation cost for high-resolution LC-ESI-MS is higher than MALDI-TOF MS, followed by the UPLC-FLR, and the cost for the CE-LIF instrumentation is the lowest. In terms of costs per sample, when the analysis is performed in high-throughput mode, both CE-LIF and MALDI-TOF MS provide low costs per sample, as they are the highest-throughput methods. In contrast, UPLC-FLD can be rather expensive, due to the low throughput of the method, and LC-ESI-MS provides the highest cost per sample as a result of the cost associated with the instrumentation as well the low throughput of the method.[77]

1.5 Summary of Subsequent Chapters

Chapter 2 describes an efficient and a simple analysis method that we developed to track the *N*-linked glycosylation sites of glycoproteins within a single experiment. The method relies on high-resolution MS data and chromatographic retention times; it identifies the glycopeptides that all bear the same peptide backbone by tracking the co-eluting incidences of those glycoforms in an LC-MS experiment using a reverse phase column. Once the approximate retention time for a particular glycosylation site is identified; the co-eluting glycopeptides at that retention time are verified with the full MS data, followed by the confirmation of the assignments by using CID data. This method was benefitted from the use of a limited *N*-linked glycan library that contained 18 glycans, which we developed by identifying abundant glycoforms in the literature of human plasma *N*-linked glycoproteins. The method successfully identified all the glycosylation sites of two model glycoproteins; thus, it was further extended to glycosylation site identification of a heavily glycosylated human plasma protein: apolipoprotein B100. The developed approach effectively mapped many of the glycosylation sites of apoB100 with a single LC-MS experiment while requiring less sample amount and less analysis time compared to the state-of-the-art analysis method.

Chapter 3 describes a rapid, direct ESI-MS approach that we developed to quantify *N*-linked glycans that ionize well in the negative ion mode. The method is straightforward; it omits glycan labeling steps; thus, reduces the need for additional post-sample clean-up steps. We successfully applied the developed (-)ESI-MS approach to quantify *N*-linked glycans of standard glycoproteins; the generated results showed higher reproducibility over multiple samples preparations and over multiple analyses. Then, the method was extended to quantify *N*-linked glycans of uromodulin: the most abundant protein excreted in human urine, which is a potential biomarker for various kidney diseases. Therefore, uromodulin was extracted directly from human urine samples of two different biological states prior to the quantitation of uromodulin *N*-linked glycan via (-)ESI-MS. The resulting *N*-linked glycosylation differences across different groups of samples were subtle; however, the method provided very tight within-group reproducibility, yielding clearly separable, unique, sample-related groups. Therefore, this method will be useful for kidney researches those use uromodulin *N*-linked glycosylation signatures to classify disease state samples, as this method is a simple and a straightforward one.

Chapter 4 describes a systematic approach used to chemically generate large sets of LC-MS glycopeptides data to mimic two different biological states; these data were generated with the goal of optimizing a newly developed supervised machine learning classifier that classifies the samples based on an entire whole glycomic profile, instead of using a single glycan feature or a few selected glycan features to classify samples. Human immunoglobulin G, IgG, is an important protein for biomarker studies; its glycosylation is affected during the progression of many diseases; thus, IgG was used in this study to generate samples of two biologically different states. We optimized the sample preparation to generate two groups of IgG glycopeptide samples, one with a native IgG glycosylation profile to mimic samples of a healthy state and the other with a slightly altered IgG glycosylation profile to mimic samples of a disease state; these samples were prepared by introducing a small percentage of partially desialylated or partially defucosylated IgG in to a native IgG tryptic digest. Finally, large sets of IgG glycopeptide data were generated for both of the samples groups; they were submitted to the newly developed Aristotle classifier and also to a more established classification approach, known as the Principal Components Analysis (PCA), to challenge their classification abilities. The generated data were challenging; yet, the Aristotle classifier outperformed the PCA while accurately classifying many of the samples of two different biological states into their accurate groups.

1.6 References

1. Cho, B. G.; Veillon, L.; Mechref, Y., *Journal of proteome research* **2019**.

2. Dalpathado, D. S.; Desaire, H., *The Analyst* **2008,** *133* (6), 731-8.

3. Zhu, Z.; Desaire, H., *Annual Review of Analytical Chemistry* **2015,** *8*, 463-483.

4. Hua, S.; An, H. J.; Ozcan, S.; Ro, G. S.; Soares, S.; DeVere-White, R.; Lebrilla, C. B., *The Analyst* **2011,** *136* (18), 3663-71.

5. Veillon, L.; Huang, Y.; Peng, W.; Dong, X.; Cho, B. G.; Mechref, Y., *Electrophoresis* **2017,** *38* (17), 2100-2114.

6. Ahuja, K. K., *The American journal of anatomy* **1985,** *174* (3), 207-23.

7. Gupta, S. K.; Bhandari, B.; Shrestha, A.; Biswal, B. K.; Palaniappan, C.; Malhotra, S. S.; Gupta, N., *Cell and tissue research* **2012,** *349* (3), 665-78.

8. Glozman, R.; Okiyoneda, T.; Mulvihill, C. M.; Rini, J. M.; Barriere, H.; Lukacs, G. L., *The Journal of cell biology* **2009,** *184* (6), 847-62.

9. Hanson, S. R.; Culyba, E. K.; Hsu, T. L.; Wong, C. H.; Kelly, J. W.; Powers, E. T., *Proc Natl Acad Sci U S A* **2009,** *106* (9), 3131-6.

10. Hang, Q.; Isaji, T.; Hou, S.; Wang, Y.; Fukuda, T.; Gu, J., *Molecular and cellular biology* **2017,** *37* (9).

11. Hsiao, C. T.; Cheng, H. W.; Huang, C. M.; Li, H. R.; Ou, M. H.; Huang, J. R.; Khoo, K. H.; Yu, H. W.; Chen, Y. Q.; Wang, Y. K.; Chiou, A.; Kuo, J. C., *Oncotarget* **2017,** *8* (41), 70653-70668.

12. Zabczynska, M.; Pochec, E., *Postepy biochemii* **2015,** *61* (2), 129-37.

13. Bosques, C. J.; Collins, B. E.; Meador, J. W., 3rd; Sarvaiya, H.; Murphy, J. L.; Dellorusso, G.; Bulik, D. A.; Hsu, I. H.; Washburn, N.; Sipsey, S. F.; Myette, J. R.; Raman, R.; Shriver, Z.; Sasisekharan, R.; Venkataraman, G., *Nature biotechnology* **2010,** *28* (11), 1153-6.

14. Shubhakar, A.; Reiding, K. R.; Gardner, R. A.; Spencer, D. I.; Fernandes, D. L.; Wuhrer, M., *Chromatographia* **2015,** *78* (5-6), 321-333.

15. Alessandri, L.; Ouellette, D.; Acquah, A.; Rieser, M.; Leblond, D.; Saltarelli, M.; Radziejewski, C.; Fujimori, T.; Correia, I., *mAbs* **2012,** *4* (4), 509-20.

16. Goetze, A. M.; Liu, Y. D.; Zhang, Z.; Shah, B.; Lee, E.; Bondarenko, P. V.; Flynn, G. C., *Glycobiology* **2011,** *21* (7), 949-59.

17. Ju, L.; Wang, Y.; Xie, Q.; Xu, X.; Li, Y.; Chen, Z.; Li, Y., *Glycobiology* **2016,** *26* (5), 460-71.

18. Kamiyama, T.; Yokoo, H.; Furukawa, J.; Kurogochi, M.; Togashi, T.; Miura, N.; Nakanishi, K.; Kamachi, H.; Kakisaka, T.; Tsuruga, Y.; Fujiyoshi, M.; Taketomi, A.; Nishimura, S.; Todo, S., *Hepatology (Baltimore, Md.)* **2013,** *57* (6), 2314-25.

19. Kyselova, Z.; Mechref, Y.; Kang, P.; Goetz, J. A.; Dobrolecki, L. E.; Sledge, G. W.; Schnaper, L.; Hickey, R. J.; Malkas, L. H.; Novotny, M. V., *Clinical chemistry* **2008,** *54* (7), 1166-75.

20. Argade, S.; Chen, T.; Shaw, T.; Berecz, Z.; Shi, W.; Choudhury, B.; Parsons, C. L.; Sur, R. L., *Urolithiasis* **2015,** *43* (4), 303-12.

21. Chen, J.; Li, X.; Edmondson, A.; Meyers, G. D.; Izumi, K.; Ackermann, A. M.; Morava, E.; Ficicioglu, C.; Bennett, M. J.; He, M., *Clinical chemistry* **2019,** *65* (5), 653-663.

22. Dong, X.; Huang, Y.; Cho, B. G.; Zhong, J.; Gautam, S.; Peng, W.; Williamson, S. D.; Banazadeh, A.; Torres-Ulloa, K. Y.; Mechref, Y., *Electrophoresis* **2018,** *39* (24), 3063-3081.

23. Liu, S.; Cheng, L.; Fu, Y.; Liu, B. F.; Liu, X., *J Proteomics* **2018,** *181*, 225-237.

24. Ruhaak, L. R.; Barkauskas, D. A.; Torres, J.; Cooke, C. L.; Wu, L. D.; Stroble, C.; Ozcan, S.; Williams, C. C.; Camorlinga, M.; Rocke, D. M.; Lebrilla, C. B.; Solnick, J. V., *EuPA open proteomics* **2015,** *6*, 1-9.

25. Ruhaak, L. R.; Taylor, S. L.; Stroble, C.; Nguyen, U. T.; Parker, E. A.; Song, T.; Lebrilla, C. B.; Rom, W. N.; Pass, H.; Kim, K.; Kelly, K.; Miyamoto, S., *Journal of proteome research* **2015,** *14* (11), 4538-49.

26. Schwedler, C.; Kaup, M.; Weiz, S.; Hoppe, M.; Braicu, E. I.; Sehouli, J.; Hoppe, B.; Tauber, R.; Berger, M.; Blanchard, V., *Analytical and bioanalytical chemistry* **2014,** *406* (28), 7185-93.

27. Su, Z.; Xie, Q.; Wang, Y.; Li, Y., *International journal of molecular sciences* **2020,** *21* (6).

28. Peng, W.; Zhao, J.; Dong, X.; Banazadeh, A.; Huang, Y.; Hussien, A.; Mechref, Y., *Expert review of proteomics* **2018,** *15* (12), 1007-1031.

29. Liu, H.; Zhang, N.; Wan, D.; Cui, M.; Liu, Z.; Liu, S., *Clinical proteomics* **2014,** *11* (1), 14.

30. Hong, Q.; Lebrilla, C. B.; Miyamoto, S.; Ruhaak, L. R., *Anal Chem* **2013,** *85* (18), 8585-93.

31. Ruhaak, L. R.; Koeleman, C. A.; Uh, H. W.; Stam, J. C.; van Heemst, D.; Maier, A. B.; Houwing-Duistermaat, J. J.; Hensbergen, P. J.; Slagboom, P. E.; Deelder, A. M.; Wuhrer, M., *PLoS One* **2013,** *8* (9), e73082.

32. Zhu, J.; Chen, Z.; Zhang, J.; An, M.; Wu, J.; Yu, Q.; Skilton, S. J.; Bern, M.; Ilker Sen, K.; Li, L.; Lubman, D. M., *Journal of proteome research* **2019,** *18* (1), 359-371.

33. Zhang, X.; Wang, Y.; Qian, Y.; Wu, X.; Zhang, Z.; Liu, X.; Zhao, R.; Zhou, L.; Ruan, Y.; Xu, J.; Liu, H.; Ren, S.; Xu, C.; Gu, J., *PLoS One* **2014,** *9* (2), e87978.

34. Hong, Q.; Ruhaak, L. R.; Stroble, C.; Parker, E.; Huang, J.; Maverakis, E.; Lebrilla, C. B., *Journal of proteome research* **2015,** *14* (12), 5179-92.

35. Kammeijer, G. S. M.; Nouta, J.; de la Rosette, J.; de Reijke, T. M.; Wuhrer, M., *Anal Chem* **2018,** *90* (7), 4414-4421.

36. Chen, I. H.; Aguilar, H. A.; Paez Paez, J. S.; Wu, X.; Pan, L.; Wendt, M. K.; Iliuk, A. B.; Zhang, Y.; Tao, W. A., *Anal Chem* **2018,** *90* (10), 6307-6313.

37. Jeong, H. J.; Kim, Y. G.; Yang, Y. H.; Kim, B. G., *Anal Chem* **2012,** *84* (7), 3453-60.

38. Reiding, K. R.; Blank, D.; Kuijper, D. M.; Deelder, A. M.; Wuhrer, M., *Anal Chem* **2014,** *86* (12), 5784-93.

39. Reiding, K. R.; Ruhaak, L. R.; Uh, H. W.; El Bouhaddani, S.; van den Akker, E. B.; Plomp, R.; McDonnell, L. A.; Houwing-Duistermaat, J. J.; Slagboom, P. E.; Beekman, M.; Wuhrer, M., *Molecular & cellular proteomics : MCP* **2017,** *16* (2), 228-242.

40. Tanabe, K.; Kitagawa, K.; Kojima, N.; Iijima, S., *Journal of proteome research* **2016,** *15* (9), 2935-44.

41. Li, Q.; Kailemia, M. J.; Merleev, A. A.; Xu, G.; Serie, D.; Danan, L. M.; Haj, F. G.; Maverakis, E.; Lebrilla, C. B., *Anal Chem* **2019,** *91* (8), 5433-5445.

42. Saldova, R.; Reuben, J. M.; Abd Hamid, U. M.; Rudd, P. M.; Cristofanilli, M., *Annals of oncology : official journal of the European Society for Medical Oncology* **2011,** *22* (5), 1113-9.

43. Reiding, K. R.; Bondt, A.; Hennig, R.; Gardner, R. A.; O'Flaherty, R.; Trbojevic-Akmacic, I.; Shubhakar, A.; Hazes, J. M. W.; Reichl, U.; Fernandes, D. L.; Pucic-Bakovic, M.; Rapp, E.; Spencer, D. I. R.; Dolhain, R.; Rudd, P. M.; Lauc, G.; Wuhrer, M., *Molecular & cellular proteomics : MCP* **2019,** *18* (1), 3-15.

44. Ruhaak, L. R.; Uh, H. W.; Beekman, M.; Hokke, C. H.; Westendorp, R. G.; Houwing-Duistermaat, J.; Wuhrer, M.; Deelder, A. M.; Slagboom, P. E., *Journal of proteome research* **2011,** *10* (4), 1667-74.

45. Wang, H.; Chen, X.; Zhang, X.; Zhang, W.; Li, Y.; Yin, H.; Shao, H.; Chen, G., *Journal of proteome research* **2016,** *15* (3), 923-32.

46. Tan, Z.; Yin, H.; Nie, S.; Lin, Z.; Zhu, J.; Ruffin, M. T.; Anderson, M. A.; Simeone, D. M.; Lubman, D. M., *Journal of proteome research* **2015,** *14* (4), 1968-78.

47. Yuan, W.; Wei, R.; Goldman, R.; Sanda, M., *Anal Chem* **2019,** *91* (14), 9206-9212.

48. Chen, B.; Zhong, X.; Feng, Y.; Snovida, S.; Xu, M.; Rogers, J.; Li, L., *Anal Chem* **2018,** *90* (2), 1129-1135.

49. Zeng, W. F.; Liu, M. Q.; Zhang, Y.; Wu, J. Q.; Fang, P.; Peng, C.; Nie, A.; Yan, G.; Cao, W.; Liu, C.; Chi, H.; Sun, R. X.; Wong, C. C.; He, S. M.; Yang, P., *Scientific reports* **2016,** *6*, 25102.

50. Shubhakar, A.; Kozak, R. P.; Reiding, K. R.; Royle, L.; Spencer, D. I.; Fernandes, D. L.; Wuhrer, M., *Anal Chem* **2016,** *88* (17), 8562-9.

51. Wei, L.; Cai, Y.; Yang, L.; Zhang, Y.; Lu, H., *Anal Chem* **2018,** *90* (17), 10442-10449.

52. Selman, M. H.; Hoffmann, M.; Zauner, G.; McDonnell, L. A.; Balog, C. I.; Rapp, E.; Deelder, A. M.; Wuhrer, M., *Proteomics* **2012,** *12* (9), 1337-48.

53. Smargiasso, N.; De Pauw, E., *Anal Chem* **2010,** *82* (22), 9248-53.

54. Miura, Y.; Hato, M.; Shinohara, Y.; Kuramoto, H.; Furukawa, J.; Kurogochi, M.; Shimaoka, H.; Tada, M.; Nakanishi, K.; Ozaki, M.; Todo, S.; Nishimura, S., *Molecular & cellular proteomics : MCP* **2008,** *7* (2), 370-7.

55. Mitra, I.; Snyder, C. M.; Zhou, X.; Campos, M. I.; Alley, W. R., Jr.; Novotny, M. V.; Jacobson, S. C., *Anal Chem* **2016,** *88* (18), 8965-71.

56. Dong, X.; Peng, W.; Yu, C. Y.; Zhou, S.; Donohoo, K. B.; Tang, H.; Mechref, Y., *Anal Chem* **2019,** *91* (18), 11794-11802.

57. Yang, S.; Wang, M.; Chen, L.; Yin, B.; Song, G.; Turko, I. V.; Phinney, K. W.; Betenbaugh, M. J.; Zhang, H.; Li, S., *Scientific reports* **2015,** *5*, 17585.

58. Zhou, S.; Hu, Y.; Veillon, L.; Snovida, S. I.; Rogers, J. C.; Saba, J.; Mechref, Y., *Anal Chem* **2016,** *88* (15), 7515-22.

59. Song, T.; Aldredge, D.; Lebrilla, C. B., *Anal Chem* **2015,** *87* (15), 7754-62.

60. Zhou, S.; Hu, Y.; DeSantos-Garcia, J. L.; Mechref, Y., *Journal of the American Society for Mass Spectrometry* **2015,** *26* (4), 596-603.

61. Tsai, T. H.; Wang, M.; Di Poto, C.; Hu, Y.; Zhou, S.; Zhao, Y.; Varghese, R. S.; Luo, Y.; Tadesse, M. G.; Ziada, D. H.; Desai, C. S.; Shetty, K.; Mechref, Y.; Ressom, H. W., *Journal of proteome research* **2014,** *13* (11), 4859-68.

62. Song, E.; Zhu, R.; Hammoud, Z. T.; Mechref, Y., *Journal of proteome research* **2014,** *13* (11), 4808-20.

63. Kim, J. Y.; Oh, D.; Kim, S. K.; Kang, D.; Moon, M. H., *Anal Chem* **2014,** *86* (15), 7650-7.

64. Miyamoto, S.; Stroble, C. D.; Taylor, S.; Hong, Q.; Lebrilla, C. B.; Leiserowitz, G. S.; Kim, K.; Ruhaak, L. R., *Journal of proteome research* **2018,** *17* (1), 222-233.

65. Shih, H. C.; Chang, M. C.; Chen, C. H.; Tsai, I. L.; Wang, S. Y.; Kuo, Y. P.; Chen, C. H.; Chang, Y. T., *Clinical proteomics* **2019,** *16*, 1.

66. Jia, X.; Chen, J.; Sun, S.; Yang, W.; Yang, S.; Shah, P.; Hoti, N.; Veltri, B.; Zhang, H., *Proteomics* **2016,** *16* (23), 2989-2996.

67. Lundstrom, S. L.; Yang, H.; Lyutvinskiy, Y.; Rutishauser, D.; Herukka, S. K.; Soininen, H.; Zubarev, R. A., *Journal of Alzheimer's disease : JAD* **2014,** *38* (3), 567-79.

68. Go, E. P.; Cupo, A.; Ringe, R.; Pugach, P.; Moore, J. P.; Desaire, H., *Journal of virology* **2016,** *90* (6), 2884-2894.

69. Go, E. P.; Ding, H.; Zhang, S.; Ringe, R. P.; Nicely, N.; Hua, D.; Steinbock, R. T.; Golabek, M.; Alin, J.; Alam, S. M., *Journal of virology* **2017,** *91* (9), e02428-16.

70. Go, E. P.; Herschhorn, A.; Gu, C.; Castillo-Menendez, L.; Zhang, S.; Mao, Y.; Chen, H.; Ding, H.; Wakefield, J. K.; Hua, D.; Liao, H.-X.; Kappes, J. C.; Sodroski, J.; Desaire, H., *Journal of virology* **2015,** *89* (16), 8245-8257.

71. Hu, W.; Su, X.; Zhu, Z.; Go, E. P.; Desaire, H., *Analytical and bioanalytical chemistry* **2017,** *409* (2), 561-570.

72. Lakbub, J. C.; Su, X.; Zhu, Z.; Patabandige, M. W.; Hua, D.; Go, E. P.; Desaire, H., *Journal of proteome research* **2017,** *16* (8), 3002-3008.

73. Ruhaak, L. R.; Kim, K.; Stroble, C.; Taylor, S. L.; Hong, Q.; Miyamoto, S.; Lebrilla, C. B.; Leiserowitz, G., *Journal of proteome research* **2016,** *15* (3), 1002-10.

74. Rebecchi, K. R.; Wenke, J. L.; Go, E. P.; Desaire, H., *Journal of the American Society for Mass Spectrometry* **2009,** *20* (6), 1048-59.

75. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., *Anal Chem* **2019,** *91* (17), 11070-11077.

76. Stockmann, H.; Adamczyk, B.; Hayes, J.; Rudd, P. M., *Anal Chem* **2013,** *85* (18), 8841-9.

77. Huffman, J. E.; Pucic-Bakovic, M.; Klaric, L.; Hennig, R.; Selman, M. H.; Vuckovic, F.; Novokmet, M.; Kristic, J.; Borowiak, M.; Muth, T.; Polasek, O.; Razdorov, G.; Gornik, O.; Plomp, R.; Theodoratou, E.; Wright, A. F.; Rudan, I.; Hayward, C.; Campbell, H.; Deelder, A. M.; Reichl, U.; Aulchenko, Y. S.; Rapp, E.; Wuhrer, M.; Lauc, G., *Molecular & cellular proteomics : MCP* **2014,** *13* (6), 1598-610.

78. Argade, S. P.; Vanichsarn, C.; Chenoweth, M.; Parsons, C. L., *BJU international* **2009,** *103* (8), 1085-9.

79. Royle, L.; Campbell, M. P.; Radcliffe, C. M.; White, D. M.; Harvey, D. J.; Abrahams, J. L.; Kim, Y. G.; Henry, G. W.; Shadick, N. A.; Weinblatt, M. E.; Lee, D. M.; Rudd, P. M.; Dwek, R. A., *Analytical biochemistry* **2008,** *376* (1), 1-12.

80. Vanderschaeghe, D.; Meuris, L.; Raes, T.; Grootaert, H.; Van Hecke, A.; Verhelst, X.; Van de Velde, F.; Lapauw, B.; Van Vlierberghe, H.; Callewaert, N., *Molecular & cellular proteomics : MCP* **2018,** *17* (12), 2508-2517.

81. Vanderschaeghe, D.; Szekrenyes, A.; Wenz, C.; Gassmann, M.; Naik, N.; Bynum, M.; Yin, H.; Delanghe, J.; Guttman, A.; Callewaert, N., *Anal Chem* **2010,** *82* (17), 7408-15.

82. Reusch, D.; Haberger, M.; Maier, B.; Maier, M.; Kloseck, R.; Zimmermann, B.; Hook, M.; Szabo, Z.; Tep, S.; Wegstein, J.; Alt, N.; Bulau, P.; Wuhrer, M., *mAbs* **2015,** *7* (1), 167-79.

83. Reusch, D.; Haberger, M.; Falck, D.; Peter, B.; Maier, B.; Gassner, J.; Hook, M.; Wagner, K.; Bonnington, L.; Bulau, P.; Wuhrer, M., *mAbs* **2015,** *7* (4), 732-42.

84. Szabo, Z.; Guttman, A.; Rejtar, T.; Karger, B. L., *Electrophoresis* **2010,** *31* (8), 1389-95.

85. Clerc, F.; Reiding, K. R.; Jansen, B. C.; Kammeijer, G. S.; Bondt, A.; Wuhrer, M., *Glycoconj J* **2016,** *33* (3), 309-43.

86. Gao, W.; Li, H.; Liu, Y.; Liu, Y.; Feng, X.; Liu, B. F.; Liu, X., *Talanta* **2016,** *161*, 554-559.

87. Parsons, C. L.; Stein, P.; Zupkas, P.; Chenoweth, M.; Argade, S. P.; Proctor, J. G.; Datta, A.; Trotter, R. N., *The Journal of urology* **2007,** *178* (6), 2665-70.

88. Chen, B.; Feng, Y.; Frost, D. C.; Zhong, X.; Buchberger, A. R.; Johnson, J.; Xu, M.; Kim, M.; Puccetti, D.; Diamond, C.; Ikonomidou, C.; Li, L., *Anal Chem* **2018,** *90* (13), 7817-7823.

89. Bladergroen, M. R.; Reiding, K. R.; Hipgrave Ederveen, A. L.; Vreeker, G. C.; Clerc, F.; Holst, S.; Bondt, A.; Wuhrer, M.; van der Burgt, Y. E., *Journal of proteome research* **2015,** *14* (9), 4080-6.

90. Stockmann, H.; O'Flaherty, R.; Adamczyk, B.; Saldova, R.; Rudd, P. M., *Integrative biology : quantitative biosciences from nano to macro* **2015,** *7* (9), 1026-32.

91. Kim, K. J.; Kim, Y. W.; Hwang, C. H.; Park, H. G.; Yang, Y. H.; Koo, M.; Kim, Y. G., *Biotechnology letters* **2015,** *37* (10), 2019-25.

92. Lauc, G.; Huffman, J. E.; Pucic, M.; Zgaga, L.; Adamczyk, B.; Muzinic, A.; Novokmet, M.; Polasek, O.; Gornik, O.; Kristic, J.; Keser, T.; Vitart, V.; Scheijen, B.; Uh, H. W.; Molokhia, M.; Patrick, A. L.; McKeigue, P.; Kolcic, I.; Lukic, I. K.; Swann, O.; van Leeuwen, F. N.; Ruhaak, L. R.; Houwing-Duistermaat, J. J.; Slagboom, P. E.; Beekman, M.; de Craen, A. J.; Deelder, A. M.; Zeng, Q.; Wang, W.; Hastie, N. D.; Gyllensten, U.; Wilson, J. F.; Wuhrer, M.; Wright, A. F.; Rudd, P. M.; Hayward, C.; Aulchenko, Y.; Campbell, H.; Rudan, I., *PLoS genetics* **2013,** *9* (1), e1003225.

93. Zhou, S.; Huang, Y.; Dong, X.; Peng, W.; Veillon, L.; Kitagawa, D. A. S.; Aquino, A. J. A.; Mechref, Y., *Anal Chem* **2017,** *89* (12), 6590-6597.

94. Peng, W.; Goli, M.; Mirzaei, P.; Mechref, Y., *Journal of proteome research* **2019,** *18* (10), 3731-3740.

95. Tao, S.; Huang, Y.; Boyes, B. E.; Orlando, R., *Anal Chem* **2014,** *86* (21), 10584-90.

96. Huang, Y.; Nie, Y.; Boyes, B.; Orlando, R., *Journal of biomolecular techniques : JBT* **2016,** *27* (3), 98-104.

97. van der Burgt, Y. E. M.; Siliakus, K. M.; Cobbaert, C. M.; Ruhaak, L. R., *Journal of proteome research* **2020**.

98. Li, H.; Gao, W.; Feng, X.; Liu, B. F.; Liu, X., *Analytica chimica acta* **2016,** *924*, 77-85.

99. Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; Kawasaki, N.; Khoo, K. H.; Kim, S.; Kondo, A.; Lattova, E.; Mechref, Y.; Miyoshi, E.; Nakamura, K.; Narimatsu, H.; Novotny, M. V.; Packer, N. H.; Perreault, H.; Peter-Katalinic, J.; Pohlentz, G.; Reinhold, V. N.; Rudd, P. M.; Suzuki, A.; Taniguchi, N., *Glycobiology* **2007,** *17* (4), 411-22.

Chapter 2 . Hunting for Hidden Glycopeptides in Highly Glycosylated Proteins

Abstract

How can you find hidden glycopeptides, which remain undetected after an LC-MS analysis of a digested glycoprotein sample? Herein we introduce a rapid and simple method for targeted identification of recalcitrant glycopeptides. A list of eighteen likely glycoforms is used to search for each hidden glycosylation site, and incidences of co-eluting peaks are investigated; multiple co-eluting peaks most often indicate the retention time of the hidden glycopeptides of interest. This strategy, of searching for co-eluting peaks, was developed because different glycoforms from the same glycosylation site tend to co-elute when subjected to reverse phase liquid chromatography. If co-eluting peaks can be found that are consistent with likely glycopeptide masses, those co-eluting peaks are putatively assigned as the hidden glycopeptides of interest. These assignments can be confirmed, and other glycoforms beyond the initial search of 18, can then be easily identified after the glycopeptides' retention time has been deciphered. We demonstrate the peak alignment approach is effective at identifying the glycosylation sites of common glycoprotein standards, and we apply it to analyze apoB100: a heavily glycosylated human plasma protein. Many glycoforms of this protein are successfully identified within a single experiment. Finally, we compared the results from the new method with a competing existing strategy, and we demonstrate that the new method is optimally effective, while requiring less sample, less instrument time, and less analysis time.

2.1 Introduction

Protein glycosylation, the covalent addition of glycans to a protein, is one of the most common post-translational modifications.[1-4] The glycans attached on proteins are crucial for regulating various biological processes, such as protein folding and stabilization,[3] cellular communication, cellular recognition and adhesion,[4] fertilization,[5] and immune defense.[6] However, aberrant glycosylation profiles of various endogenous proteins have been associated with the progression of various diseases such as cancers,[7-9] congenital disorders,[10] and inflammatory diseases.[11-12] Thus, glycosylation changes on some proteins aid as prognostic biomarkers for various disease states,[13-14] while some glycosylated proteins, such as the HIV-1 envelope glycoprotein, (Env,) [15-16] serve as a major target for vaccine development. Therefore, development of effective protein glycosylation site profiling strategies is important to identify the alterations of site-specific glycosylation during the progression of various diseases and also to develop effective vaccine candidates and biotherapeutics.

In recent years, mass spectrometry (MS) has emerged as the most powerful technique for glycosylation analysis because of its high resolution, high sensitivity, and the availability of complementary tandem mass spectrometry (MS/MS) techniques.[15, 17-18] A commonly employed workflow for glycopeptide-based analysis using mass spectrometry includes proteolytic digestion of purified glycoproteins, HPLC separation, and detection using MS, followed by data analysis with software tools for glycopeptide identification, which attempt to assign all possible glycopeptides present.[19] Recently, many bioinformatics tools, such as Byonic,[20-21] GlycoPep Grader (GPG),[22] MAGIC[23] and Glycopeptide search[24] have been developed to improve the *N*-linked glycopeptide identification. In general, these bioinformatics tools assign the best-matching glycopeptide composition to a CID or ETD spectrum by identifying glycopeptides' specific fragment ions in the resultant MS/MS data. The accurate identification and assignment of glycopeptides critically depends on the quality of the resultant MS/MS data, so the tools are biased towards detecting particular glycopeptides that ionize well and have easily assignable MS/MS data.[25] The current tools are inadequate when an exhaustive analysis is required, since they often only detect the easy-to-find glycoforms. For example, MAGIC only identified 36 glycopeptides from the HeLa cell proteome.[23]

When automated, untargeted strategies fail to identify all the glycoforms present in a sample, or when more complete glycosylation coverage is required, expert targeted analysis approaches are used, where additional experiments are conducted and MS and MS/MS data are manually interpreted to provide more complete glycosylation coverage. The benefits of targeted, expert analysis can be dramatic: in one example, two different labs characterized the same protein using either expert analysis or an automated approach that relied on Bionic software: the experts identified ~600 glycopeptides,[26] while the Bionic-based approach identified ~160 glycopeptides.[27] While the benefits of expert analysis can clearly be seen in increased coverage, the cost in time and expertise is massive. Targeted glycopeptide analysis strategies, which can be done expediently by relative novices, therefore, would greatly benefit the field.

One strategy to find reclusive glycosylation sites is shown in Figure 1A. The method is tedious but reasonably effective. PNGase F is used to deglycosylate the glycoprotein followed by two parallel LC-MS analyses: the first relies on the analysis of deglycosylated peptides to identify the approximate retention time(s) of the peptides of interest; the second analysis, of glycosylated peptides using the same LC-MS conditions, identifies the glycoforms that correspond to the glycopeptides eluting at the retention time(s) of interest.[25, 28] Many times this approach is not optimal, due to the need for increased sample amounts. Additionally, the approach requires at least twice the sample preparation and data analysis time, since two different LC-MS files need to be run and interpreted.

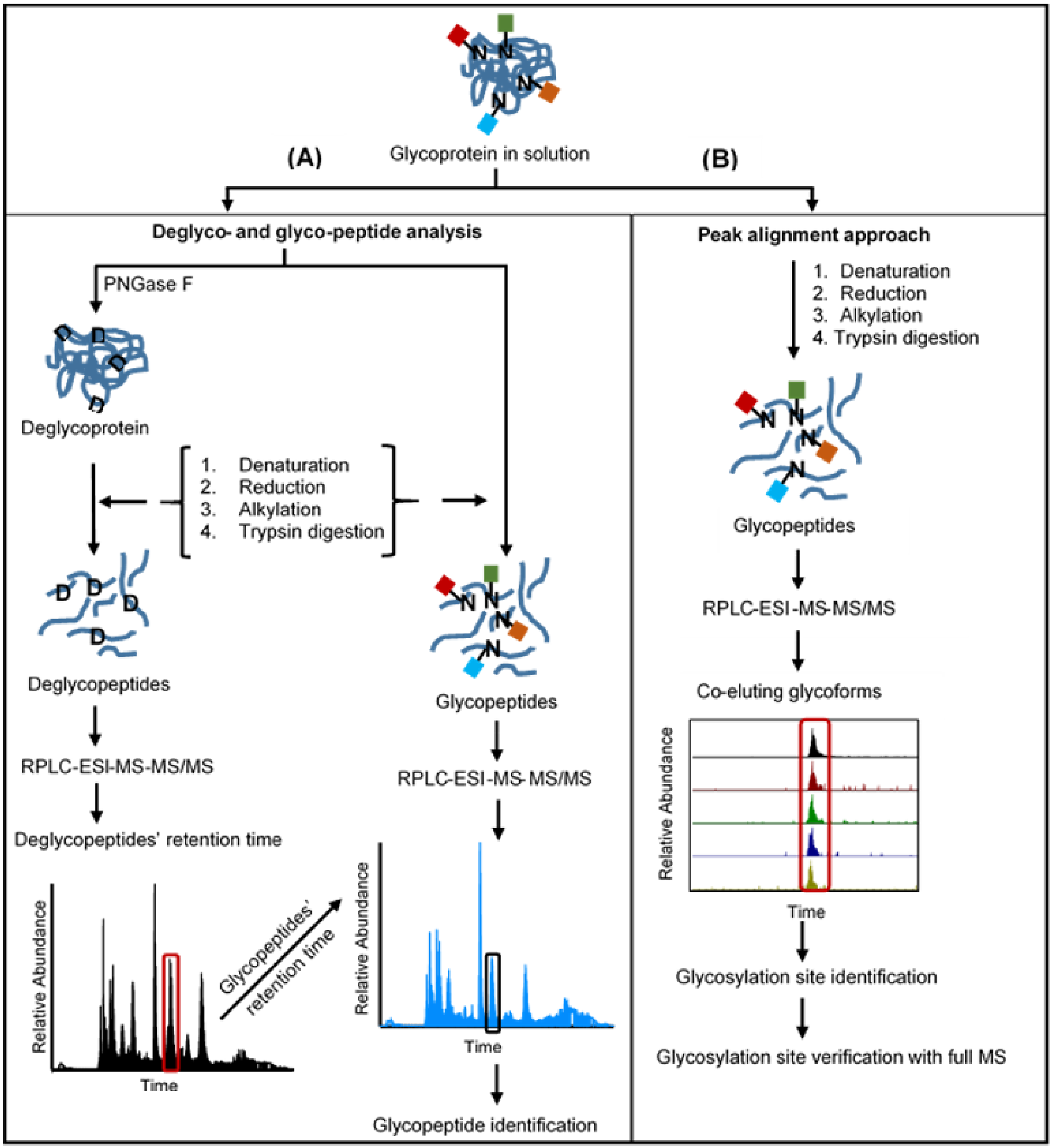


Figure 1. Schematic representation of the experimental workflow for *N*-linked glycosylation site mapping using the existing approach (deglyco- and glyco-peptide analysis) **(A)** and novel peak alignment approach **(B)**.

Herein, we introduce an efficient and simple strategy that identifies hidden *N*-linked glycopeptides, even if they ionized poorly and/or if they generate sub-optimal MS/MS data that could not be effectively assigned in an automated, untargeted search. The novel strategy depends on using high-resolution MS data and chromatographic retention times to map the *N*-linked glycosylation sites of the protein, as shown in Figure 1B. When using the new method, one searches for a set of common or likely glycopeptides that all contain the same peptide backbone; if they are present, they all co-elute on a reverse phase column. Therefore, when peaks corresponding to the masses of different glycopeptides appear at a similar retention time, one can provisionally assign those as glycopeptides for a particular glycosylation site and confirm with the available high-resolution MS and MS/MS data. The method was validated on two model glycoprotein standards and extended to glycosylation sites analysis of apoB100, a heavily glycosylated human plasma glycoprotein. Finally, we compared the results of the novel method with an existing targeted analysis strategy and verified that our method is optimally effective, while requiring half the sample prep, half the sample quantity, half the instrument time, and less analysis time.

2.2 Experimental

2.2.1 Materials and Reagents

Bovine ribonuclease B (RNAse B), bovine fetuin and apolipoprotein B from human plasma were purchased from Sigma Aldrich (St. Louis, MO). Sequencing grade trypsin was acquired from Promega (Madison, WI), and PNGase F (500, 000 units/mL) was from New England BioLabs (Ipswich, MA). All the chemical reagents were of analytical grade or better.

2.2.2 Glycoprotein Digestion

The glycoproteins (100 – 250 µg) were dissolved in 100 mM Tris-HCl buffer (pH 8.5) to give ≥ 4 mg/mL initial concentration; they were thermally denatured by at 100 °C for 10 mins. The samples were cooled to room temperature followed by the addition of urea to a final concentration of 6 M for further denaturation. For reduction and alkylation of disulfide bonds, tris (2-carboxyethyl)-phosphine (TCEP) and iodoacetamide (IAM) were added to give a final sample concentration of 5 mM and 25 mM, respectively. The samples were then incubated for 1 h at room temperature in the dark. Subsequently, the alkylation reaction was quenched via the addition of dithiothreitol (DTT) to give a final concentration of 30 mM followed by a 30 min incubation period at room temperature. For the apoB100, TCEP and IAM were added to give a final sample concentration of 0.6 mM and 1 mM, followed by the neutralization of excess IAM via the addition of DTT (2 mM final concentration). After these steps, samples were diluted with Tris- HCl buffer (pH 8.5) to a final concentration of 1 µg/µL, followed by the addition of trypsin at an enzyme-to-protein ratio of 1:30 (w/w) and an incubation period of 18 h at 37 ˚C. This step was followed by a second trypsin addition at 1:100 enzyme-to-protein ratio (w/w) to ensure complete digestion of the protein samples. The trypsin digestion of all the samples were then quenched by adding 1 µL of acetic acid for every 100 µL of sample. All the digested solutions were stored at -20 ˚C until the analysis, except the apoB100 sample, which required an additional step of centrifugation at 10,000 g for 4 min to obtain the supernatant, prior to the storage.

2.2.3 Glycoprotein Deglycosylation

Glycoprotein samples in Tris-HCl buffer (protein concentration of ≥4 mg/mL) at pH 8.5, were thermally denatured at 100 ˚C for 10 min and cooled to room temperature. To carry out the deglycosylation, for RNAse B, fetuin, and apoB100, 1 µL, 2 µL, and 6 µL of PNGase F stock solution (50 units /µL solution in H_2O) was added. The first two samples were incubated at 37 °C overnight while the latter was incubated for 24 h, at pH 8.5. The deglycosylated samples were then subjected to trypsin digestion by following the same experimental conditions as described in the glycoprotein digestion section. The prepared solutions were stored at -20 °C prior to the analysis.

2.2.4 LC-MS Analysis

LC-MS analysis of digested glycoprotein samples were performed on an LTQ Orbitrap Velos Pro hybrid mass spectrometer (Thermo Scientific, San Jose, CA) coupled to an Acquity UPLC system (Waters, Milford, MA). Digested glycoprotein samples; 5 µL of RNAse B (5 µM), fetuin (1 µM) or apoB100 (1 µM) were injected on to a C_{18} column (320 µm i.d. × 5 cm, 3.5 µm pore size, Micro-Tech, Vista, CA) at a flow rate of 10 µL/min, for the separation. Mobile Phase A consists of 99.9% H_2O plus 0.1% formic acid, while Mobile Phase B consists of 99.9% CH_3CN plus 0.1% formic acid. Different gradients were used to maximize the glycopeptide separation of different glycoprotein samples. For RNAse B, 2% Mobile Phase B for 5 min, 5% to 15% B in 5 min, and 15% to 35% linear increase of B in next 40 min was used. For fetuin and apoB100, 2% B for 5 min, 2% to 12% B in 2 min, 12% to 35% linear increase of B in 43 mins were used. After the separation gradient of each glycoprotein digest, the column was washed by increasing Mobile Phase B to 80% in 10 min and holding it at 80% B for another 10 min followed by decreasing the B to 5% in 5 min and subsequent re-equilibrating the column at 2% B for another 10 min. The LC-MS analysis of deglycosylated and corresponding glycosylated protein digests were run back-to-back with a wash and blank run in between to minimize the sample carry over.

For mass spectrometric analysis, positive ion mode was used with an ESI spray voltage of 3.0 kV and capillary temperature of 260 ˚C (250 ˚C for apoB100). For all the experiments, full scan mass spectra were acquired in the Orbitrap, for the mass range of (*m/z* 400 – 2000) at a resolution of 30,000 (at *m/z* 400). The CID data were acquired in a data dependent mode to confirm the compositional assignments of glycopeptides. The CID spectra were collected in the ion trap by picking the top 8 intense ions (top 5 for apoB100) of the full MS, with a repeat count of one, repeat duration of 30 s, and dynamic exclusion window of 180 s. For each CID scan, normalized collision energy of 35%, and an activation time of 10 ms was used.

2.3 Results and Discussion

2.3.1 Method Conception and Overview

One important principal of the method described here is that different glycoforms of the same glycosylated peptide typically co-elute when subjected to RP-HPLC. Figure 2A demonstrates this by representing the extracted ion chromatograms (XICs) of multiple doubly charged glycoforms of the **N^{33}**LTK glycopeptide, generated from tryptic digest of RNAse B. As shown in this figure, the glycoforms co-eluted within the retention time range of (1.88 – 3.97) min. Using this principal, that on a reverse-phase column, the glycoforms would co-elute while allowing the identification of likely retention time range for a set of glycoforms of interest, which can be verified by using high resolution MS data as shown in Figure 2B. Thus, we developed a method to identify the retention time of any new glycopepetide by searching for glycoforms that were likely to be present in the sample and finding instances of co-eluting peaks in the XICs.

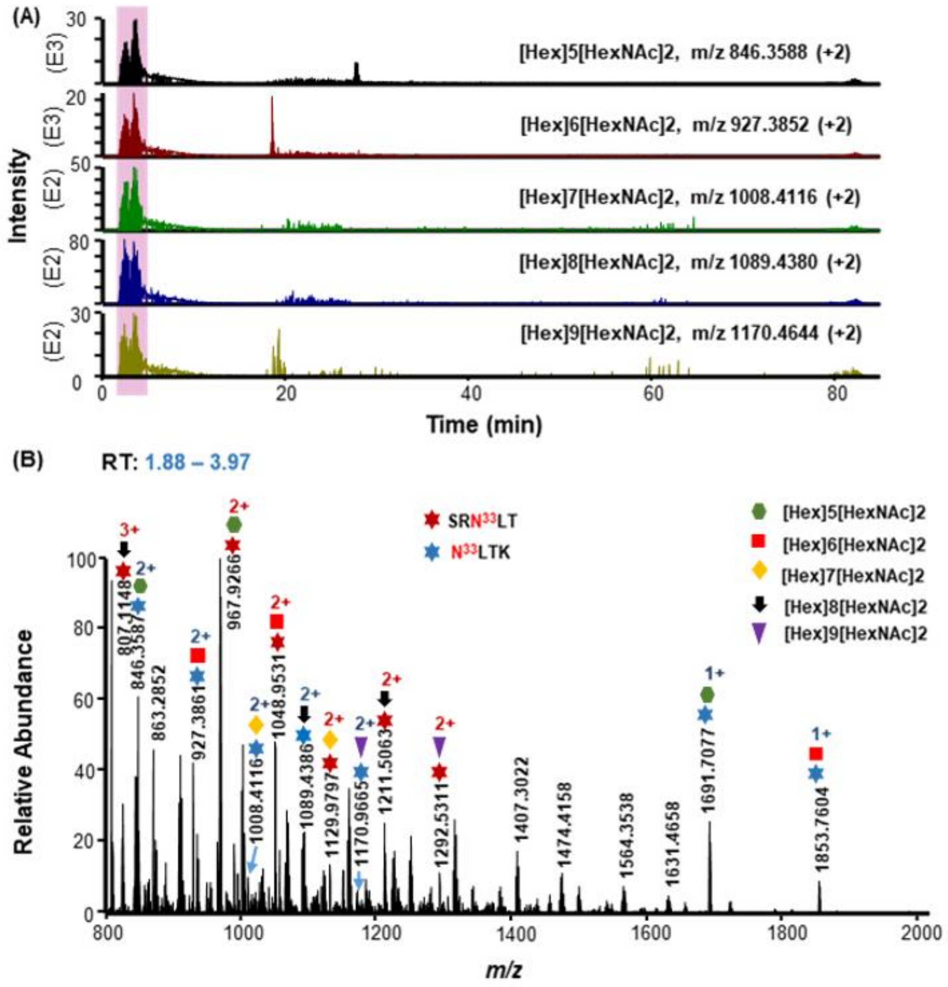


Figure 2. The XICs' for the doubly charged multiple glycoforms of the **N^{33}**LTK glycopeptide of RNAse B, co-eluted within the retention time range of (1.88 – 3.97) min **(A)**. High resolution MS spectrum recorded for the tryptic digest of RNAse B for the time range of (1.88 – 3.97) mins **(B)**. This full MS shows two co-eluting tryptic glycopeptides of RNAse B, a missed cleaved SR**N^{33}**LTK (red star) and **N^{33}**LTK (blue star) glycopeptides. Each glycopeptide was denoted with different symbols (as shown in the legend) to represent various high-mannose type glycans attached to the glycosylation site.

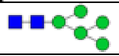
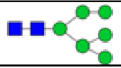
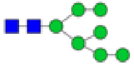
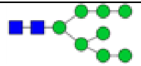
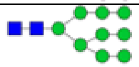
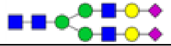
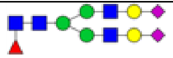
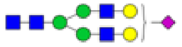
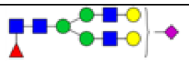
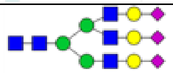
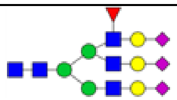
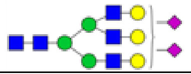
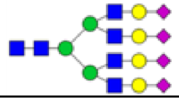
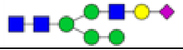
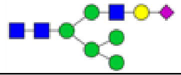
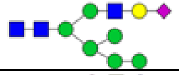
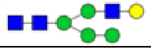
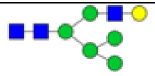
The key intellectual challenge in developing this method is determining which glycoforms to search for when the glycosylation profile is unknown. To solve this problem, we developed a representative *N*-linked glycan library, using the most common literature-reported glycan compositions for the human plasma glycoproteins.[29-31] We selected 18 of the most abundant glycoforms (shown in Table 1), and use those to search the XICs and identify incidences of co-eluting peaks. After the glycosylation site is tentatively identified by the presence of co-eluting peaks corresponding to possible glycoforms, the glycopeptides in that portion of the chromatogram are assigned based on high-resolution mass and further confirmed by manually interpreting corresponding CID data.

2.3.2 Developing the *N*-linked Glycan Library

The success of this method hinges on identifying a list of glycans, where at least a few of them can be reasonably expected to be present on the glycopeptides. To build the limited glycan library for this workflow, we used literature reported information of human plasma proteins' *N*-linked glycosylation.[29-31] Glycosylation information reported for 20 different human plasma glycoproteins, including Alpha-1-acid glycoprotein, Alpha-1-antitrypsin, Immunoglobulins, Ceruloplasmin, Fibrinogen, Vitronectin, and others were used. The identified glycan compositions for each glycoprotein were compiled for each reported glycosylation site on the proteins. The glycoforms were tracked in three major categories: high mannose, complex, and hybrid. All the instances of all the different glycans detected were tabulated; in total, 225 glycans on 83 different glycosylation sites were recorded; See Appendix A, Table 1. Using these data, representative *N*-linked glycans were chosen to build the limited glycan library by considering the number of times each glycan appears on different proteins and the total number of proteins that contain a particular glycan composition. By doing so, we identified the most commonly introduced *N*-linked glycans for abundant human plasma glycoproteins and picked 18 glycan compositions to use in a limited glycan library. These glycans were selected to cover all three major glycans types (high-mannose, complex, and hybrid); therefore, five, eight and, five

N-linked glycan compositions were included for high mannose, complex and hybrid groups, respectively (see Table 1).

Table 1. The developed *N*-linked glycan library for human plasma glycoproteins. "A" denotes for number of antennae.

Group	Glycan Composition	Glycan Structure	Glycan Mass/Da
1	**High Mannose**		
M5	[Hex]5[HexNAc]2		1234.4334
M6	[Hex]6[HexNAc]2		1396.4862
M7	[Hex]7[HexNAc]2		1558.5390
M8	[Hex]8[HexNAc]2		1720.5918
M9	[Hex]9[HexNAc]2		1882.6446
2	**Complex**		
A2G2S2	[Hex]5[HexNAc]4[NeuAc]2		2222.7830
FA2G2S2	[Hex]5[HexNAc]4[Fuc]1[NeuAc]2		2368.8409
A2G2S1	[Hex]5[HexNAc]4[NeuAc]1		1931.6876
FA2G2S1	[Hex]5[HexNAc]4[Fuc]1[NeuAc]1		2077.7455
A3G3S3	[Hex]6[HexNAc]5[NeuAc]3		2879.0106
A3FG3S3	[Hex]6[HexNAc]5[Fuc]1[NeuAc]3		3025.0685
A3G3S2	[Hex]6[HexNAc]5[NeuAc]2		2587.9152
A4G4S4	[Hex]7[HexNAc]6[NeuAc]4		3535.2382
3	**Hybrid**		
M4G1S1	[Hex]5[HexNAc]3[NeuAc]1		1728.6082
M5G1S1	[Hex]6[HexNAc]3[NeuAc]1		1890.6610
M6G1S1	[Hex]7[HexNAc]3[NeuAc]1		2052.7138
M4G1	[Hex]5[HexNAc]3		1437.5128
M5G1	[Hex]6[HexNAc]3		1599.5656

Our rationale for choosing 18 glycans, and not more or less, is somewhat arbitrary: clearly, searching with a larger library might be advantageous in some cases; using a larger library could be helpful if the limited library proves insufficient for a given analysis. Choosing fewer than 18 glycans would nominally speed up the search, since fewer chromatograms would need to be extracted. Balancing these competing considerations, we chose a relatively small set of abundant glycoforms and determined the utility of this limited library on several proteins. All of the 18 *N*-linked glycans compositions in the library were used together for the targeted identification of particular glycosylation sites in several proteins by searching co-eluting glycoforms from a reverse phase column.

2.3.3 Using the Library to Search for Co-eluting Peaks

Monoisotopic glycoform masses for a glycopeptide of interests are generated by adding the monoisotopic peptide mass to that of the glycan masses in the library. Secondly, the *m/z*'s for the multiple charge states of those glycoforms are generated. Next, the 18 glycoforms' *m/z*'s that bear the least possible charge state within the instrument scan range (*m/z* 400 – 2000) are considered first. In here, a single charge state of eighteen glycoforms of a particular peptide is taken into account, if all of the glycoform *m/z*'s of that specific charge state lie within the instrument scan range. If not, the *m/z*'s of two consecutive charge states of glycoforms are combined to generate 18 *m/z*'s for the extraction. For example, during the glycosylation site search of fetuin, two consecutive *m/z*'s of RPTGEVYDIEIDTLETTCHVLDPTPLA**N**$^{\mathbf{81}}$CSVR (3+ and 4+ charge states) and LCPDCPLLAPL**N**$^{\mathbf{138}}$DSR (2+ and 3+ charge states) glycopeptides were considered together for the extraction, as shown in Figures 3A and 3B. At the end of the glycoform extraction, the retention time of the glycopeptide of interest is provisionally identified by the presence of co-eluting peaks in the XIC's. Once the retention time is found, that region of the chromatogram can easily be manually investigated to determine all glycoforms present at that site. Examples are included below.

2.3.4 Method Demonstration

The method was applied to a well-characterized model glycoprotein; fetuin. Figure 3A and 3B represent the resultant extracted ion chromatograms obtained for the 18 glycoforms of fetuin peptides RPTGEVYDIEIDTLETTCHVLDPTPLA**N**$^{\mathbf{81}}$CSVR and LCPDCPLLAPL**N**$^{\mathbf{138}}$DSR, respectively. In these figures, a set of co-eluting peaks corresponding to the two glycopeptides mentioned above were identified in the time ranges of 39.00 – 41.00 min (Figure 3A) and 25.00 – 27.00 min (Figure 3B). This result allowed the rapid identification of N81 and N138 glycosylation sites respectively. Afterwards, the provisionally assigned glycosylation sites were verified with the full MS data generated for the retention time ranges where the co-eluting peaks were identified. Figure 3C and Figure 3D illustrate the high-resolution MS data generated for the N81 and N138 glycosylation sites, respectively.

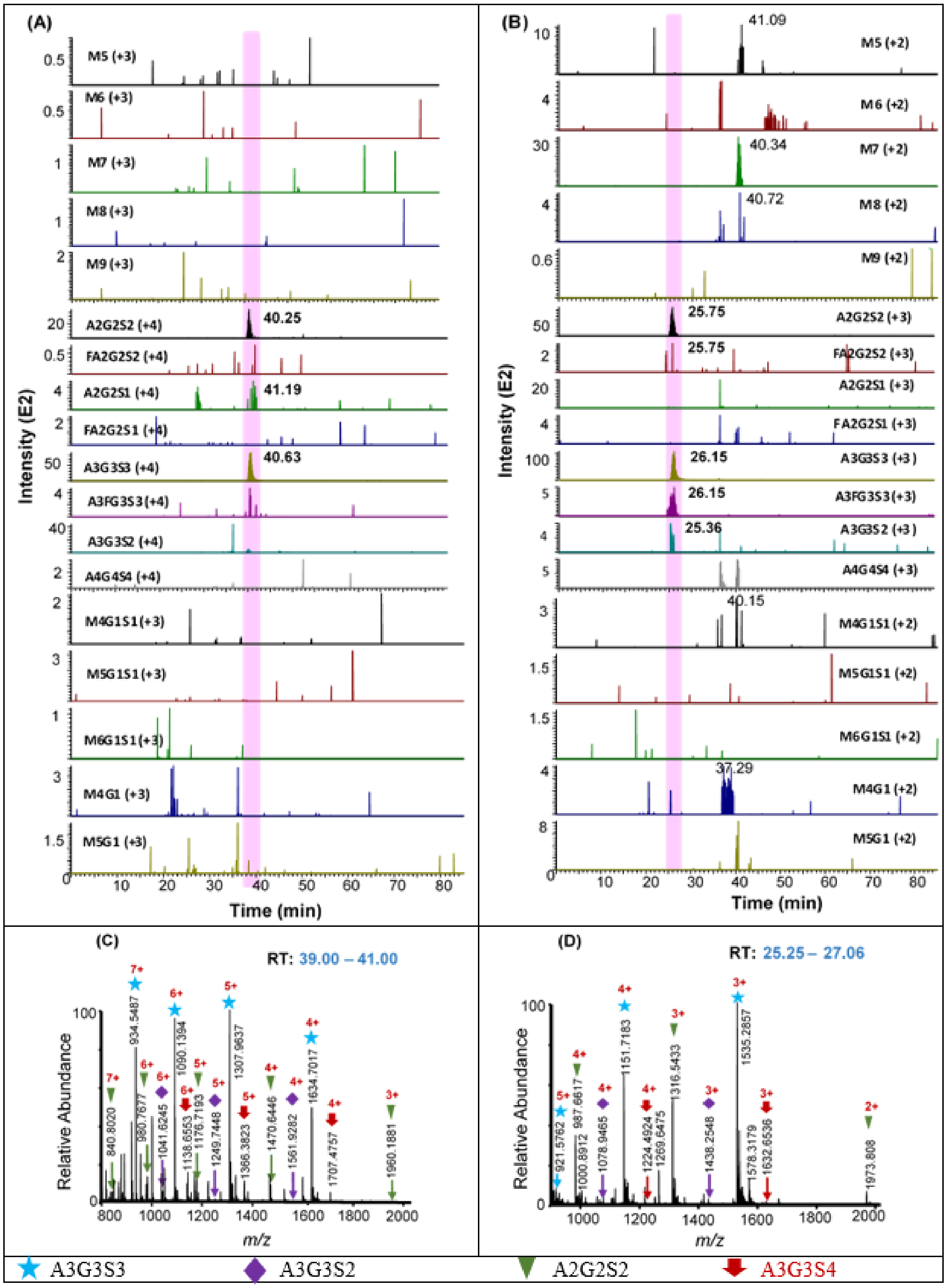


Figure 3. The XIC's of eighteen glycoforms of two fetuin glycopeptides, RPTGEVYDIEIDTLETTCHV LDPTPLAN^{81}CSVR at +3 and +4 charge states **(A)** and LCPDCPLLAPLN^{138}DSR at +2 and +3 charge states **(B)**. The corresponding high-resolution MS spectra of the N81 **(C)** and N138 **(D)** sites containing glycopeptides recorded for the time ranges of (39.00 – 41.00) min and (25.25 – 27.06) min. A set of co-eluting glycoforms (A2G2S2, A3G3S3 and A3G3S2) were identified for both N81 and N138 glycosylation sites and the co-eluting time ranges are highlighted in pink.

In both of these figures, glycoforms corresponding to multiple charge states of A2G2S2, A3G3S3, and A3G3S2 were identified, verifying the rapid glycosylation site assignment of N81 and N138 glycosylation sites. Finally, these glycopeptide assignments were further confirmed by manually validating the corresponding CID data. Once the glycosylation site is assigned, the rest of the candidate glycoform compositions that all contain the same glycosylation site, but are not included in the *N*-linked glycan library, are identified. For example, the parent ion *m/z'*s can be transported to Glycomod, and the resulting glycoform candidates can be validated with available CID data. By using this strategy, an additional glycoform: [Hex]6[HexNAc]5[NeuAc]4 (A3G3S4) was confirmed for both N81 and N138 glycosylation sites of fetuin. In addition to the aforementioned glycosylation sites, we applied the method to identify glycopeptides of N158 glycosylation site; the third glycosylation site of fetuin that bears the VVHAVEVALATFNAES**N**$^{\mathbf{158}}$GSYLQLVEISR peptide backbone. Figure 4 shows the XIC's of eighteen glycoforms of fetuin glycopeptide: VVHAVEVALATFNAES**N**$^{\mathbf{158}}$GSYLQLVEISR at +4 charge state.

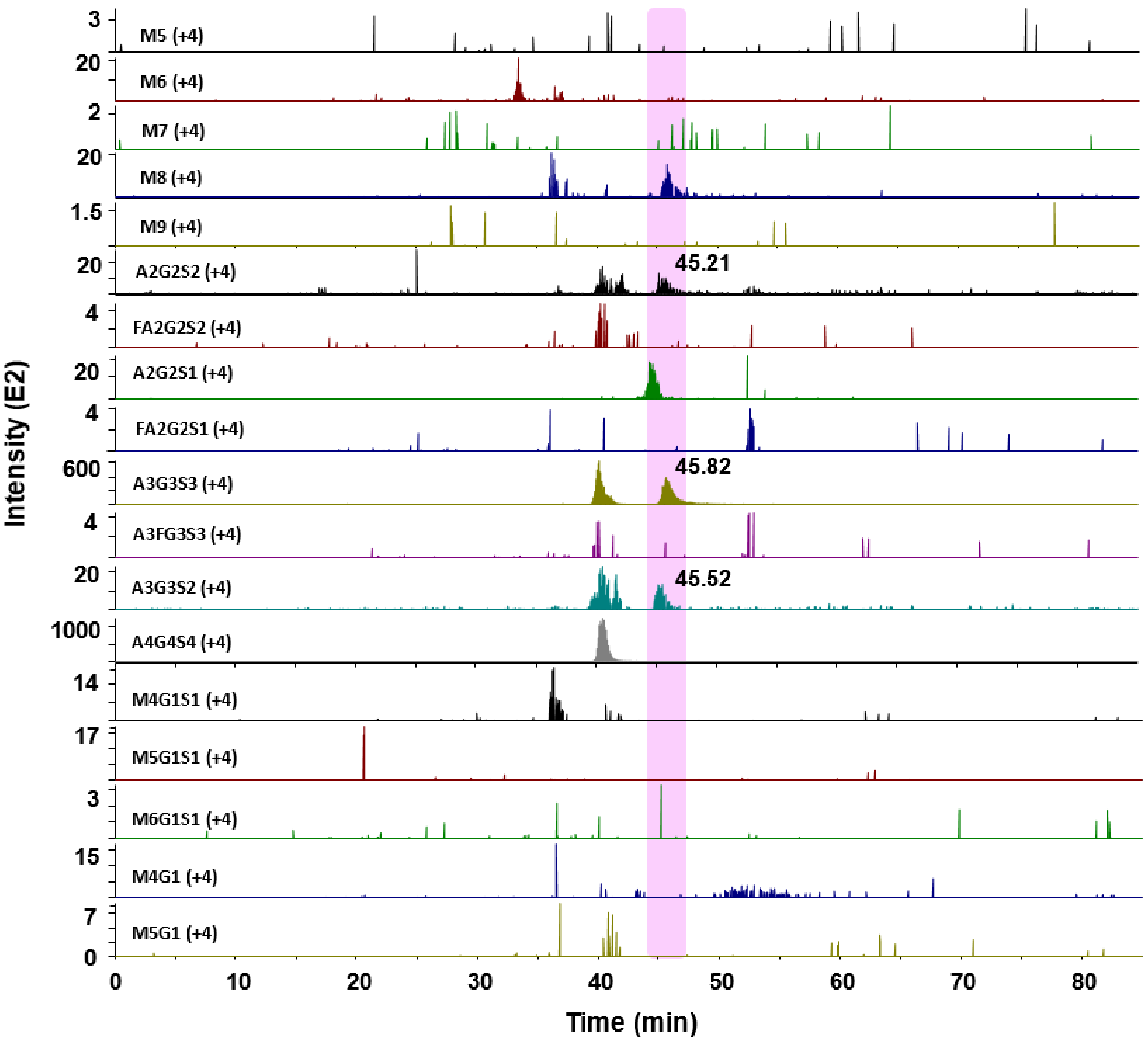


Figure 4. The XIC's of eighteen glycoforms of fetuin glycopeptide, VVHAVEVALATFNAES**N**^{158}GSY LQLVEISR at +4 charge state. A set of co-eluting glycoforms (A2G2S2, A3G3S3 and A3G3S2) were identified for N158 glycosylation site at the retention time range of (45.00 – 47.00) min. The co-eluting time range is highlighted in pink.

Similar to the N81 and N138 glycosylation sites of fetuin, the method afforded the identification of A2G2S2, A3G3S3 and A3G3S2 from the library and A3G3S4, which was not included in the limited glycan library. Finally, multiple charge states of these glycoforms were identified, verified, and confirmed with the full MS and CID data, as described earlier.

Users of this peak alignment approach should be aware that sometimes more than one set of co-eluting peaks are observed in the extracted ion chromatograms. For example, two sets of co-eluting peaks were observed for N158 glycosylation site of fetuin (Figure 4) at (39.00 – 41.00) min and (45.00 – 47.00) min, respectively. In such instances, the correct set of co-eluting peaks can be easily identified by querying the high-resolution MS data at both retention times. In the set of peaks eluting at ~45 minutes, the extracted peaks indeed matched the theoretical monoisotopic masses of the glycopeptides that were extracted. This was not the case for the peaks eluting around 40 minutes. Here, the extracted peaks' monoisotopic, high-resolution masses were not consistent with the masses of the glycopeptide ions that were searched. In addition to verification by high-resolution mass, the correct co-eluting peaks are also verified by assigning CID data for the peaks.

2.3.5 Method Validation

To validate the glycosylation site identification results obtained for fetuin with the novel approach, we used an established analysis method, which is shown in Figure 1A. It relies on two parallel LC-MS experiments. In this workflow, the retention time for a particular deglycosylated tryptic peptide of interest is identified in a preliminary LC-MS experiment, and glycopeptides are identified in a second LC-MS file by searching at the same retention times. The glycopeptides in this retention time region are putatively assigned by the full MS data, and they are confirmed with CID data. Figure 5 shows an example of the two parallel LC-MS experiments for the identification of N81 glycosylation site of fetuin.

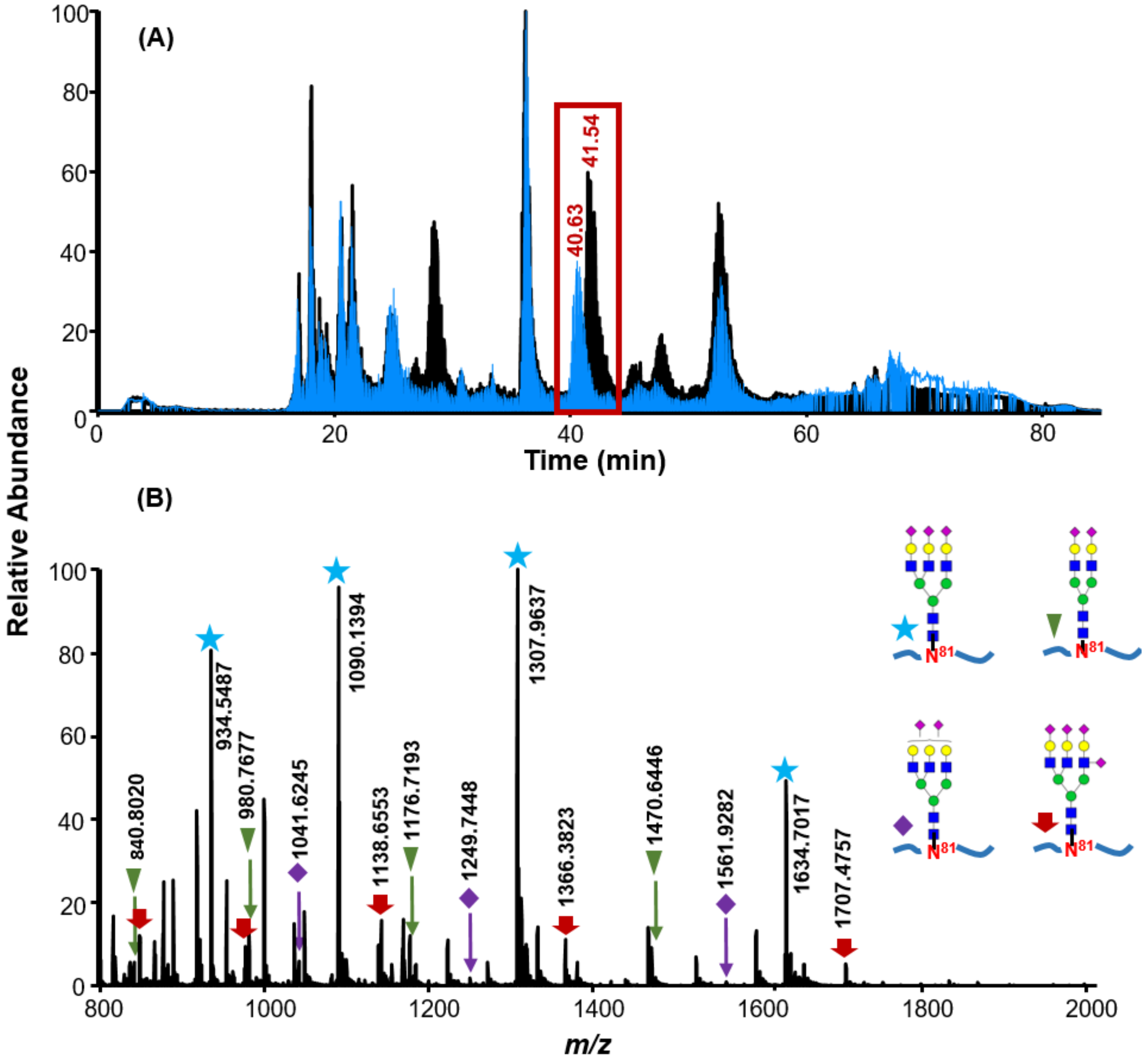


Figure 5. The overlapped total ion chromatograms of the tryptic digests of deglycosylated (black) and glycosylated (blue) fetuin samples ran under the same LC-MS conditions **(A)**. The highlighted time range (red box) shows the retention times of fetuin deglycopeptide (black color peak, 41.54 min) and corresponding glycopeptide (blue color peak, 40.63 min) that each contains N81 glycosylation site. The full MS generated within the retention time range of (39.00 – 41.00) min for the fetuin glycopeptide (blue color peak) that contains N81 glycosylation site **(B)**. The symbols represent the peaks correspond to the different glycoforms of N81 glycosylation site at their multiple charge states, A3G3S3 (five point blue star), A3G3S2 (purple diamond), A2G2S2 (green triangle), and A3G3S4 (red arrow).

In Figure 5A, deglycosylated RPTGEVYDIEIDTLETTCHVLDPTPLA**D^{81}**CSVR peptide elutes at 40.00 – 43.00 min, while the corresponding glycosylated peptide elutes at 39.00 – 42.00 min. Figure 5B shows the high-resolution mass spectrum at this retention time, which includes the RPTGEVYDIEIDTLETTCHVLDPTPLA**N^{81}**CSVR glycopeptides. Four different glycoforms: A3G3S3, A3G3S2, A2G2S2, and [Hex]6[HexNAc]5[NeuAc]4 (A3G3S4) are identified in multiple charge states. The glycoforms identified in this approach for the N81 glycosylation site of fetuin are consistent with the literature data.[32] Importantly, both literature precedent and this existing approach give identical results to the new approach we developed, which does not require analysis of a deglycosylated sample.

Since we were able to use the developed glycan library along with the full MS data to successfully track the glycosylation sites and all the glycoforms of both RNAse B and fetuin, we applied the method to a more complex human plasma protein: apolipoprotein B100.

2.3.6 Apolipoprotein B100 Demonstration

Apolipoprotein B100 is a 550 kDa protein with 19 potential *N*-linked glycosylation sites. It is the major protein component in low-density lipoprotein (LDL) and is recognized by the LDL receptor. A detailed characterization of the site-specific *N*-linked glycosylation of apoB100 has been reported previously.[29] This protein is a suitable candidate for the assessment of the novel peak alignment approach because of its biological importance and its significant number of glycosylation sites.

Hence, apolipoprotein B100 was used to test the efficiency of the developed *N*-linked glycan library for identifying glycosylation sites of complex glycoproteins. The same 18 glycoforms selected previously were used to identify the retention time of the glycosylation sites by checking for the incidences of co-eluting peaks. Once the retention time is identified, the glycosylation sites were provisionally assigned; they were verified and confirmed with high-resolution MS and CID data, respectively.

Figure 6 shows an example where this approach successfully identifies the retention time and glycan type for the N158 (G2) glycosylation site of apoB100.

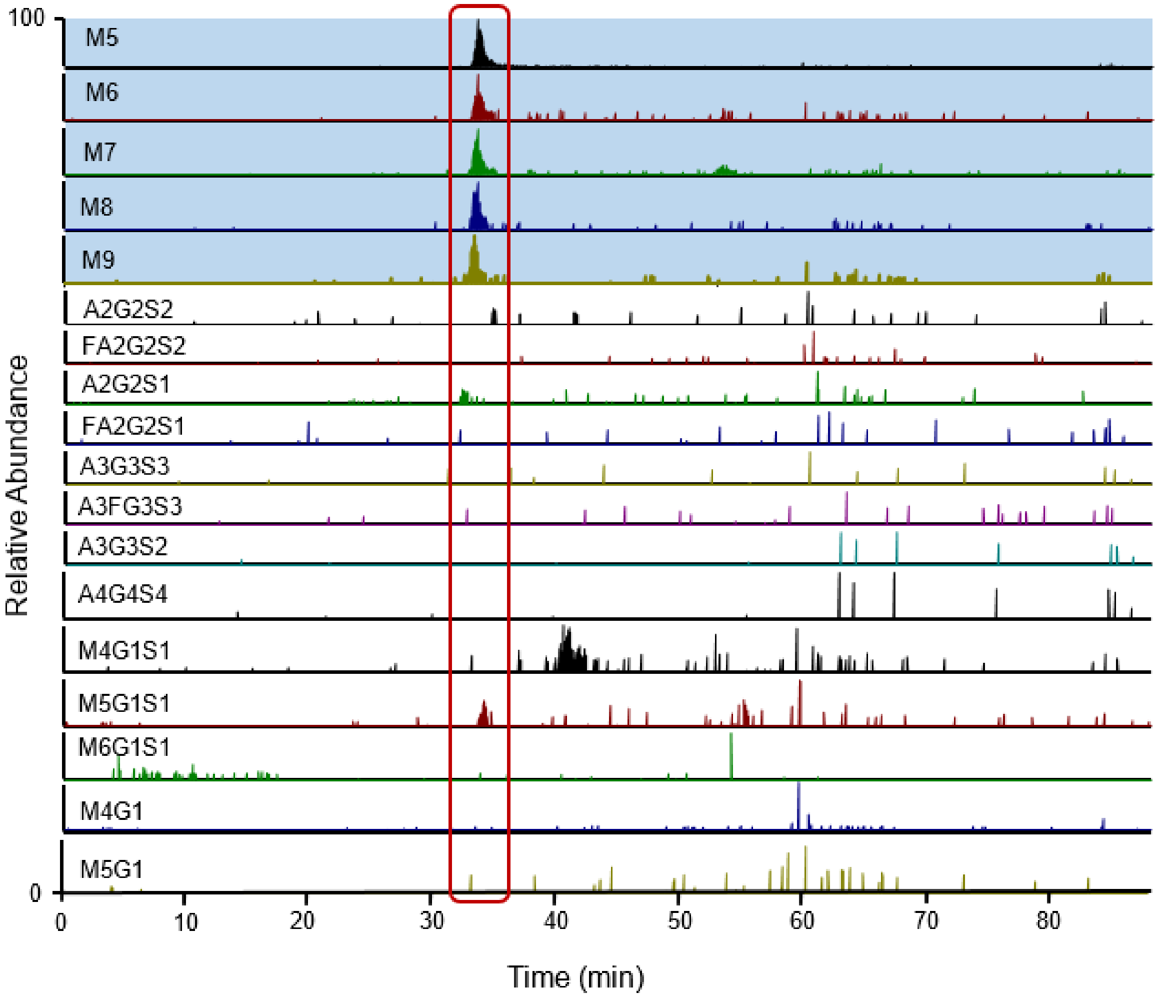


Figure 6. Representative data showing the application of the peak alignment approach to identify the N158 (G2) glycosylation site of apoB100. The XIC's of 18 triply-charged glycoforms obtained for the QVLFLDTVYG**N**[158]CSTHFTVK glycopeptide with corresponds to the *N*-linked glycan library: the co-elution of five high-mannose type glycans (M5 to M9), highlighted in blue, were observed in the time range of (31.50 – 34.50) mins. The *N*-linked glycosylation site identification was confirmed by full MS and MS/MS data. For details on all of the identified glycoforms, see Appendix A, Table 2.

The *m/z*'s corresponding to eighteen glycoforms generated for the QVLFLDTVYG**N**$^{\mathbf{158}}$CSTHFTVK glycopeptide at 3+ charge state were extracted, and the resultant XIC's were aligned to identify co-eluting glycoforms. The generated XIC's of the G2 glycosylation site show a clear co-elution of a set of high-mannose glycoforms within the time range of (31.50 – 34.50) min, indicating the retention time for the QVLFLDTVYG**N**$^{\mathbf{158}}$CSTHFTVK peptide. The full MS data generated for the retention time of interest (31.50 – 34.50) min were searched to further verify the glycoforms on this peptide. Appendix A, Table 2 represent the *N*-linked glycopeptides of apoB100 extracted and verified with the full MS and MS/MS data. These data confirmed that the G2 glycosylation site is occupied by high-mannose glycans. The resultant glycoform identification data were then compared with the literature reported information,[29] to assess the reliability of the new strategy. The new approach successfully identified the glycosylation site and all the glycoforms (M5-M9) reported for the N158 glycosylation site of apoB100, consistent with the literature.

Figure 7 shows another example where the workflow described here clearly identified the N1496 (G6) glycosylation site; this site has a higher glycan diversity.

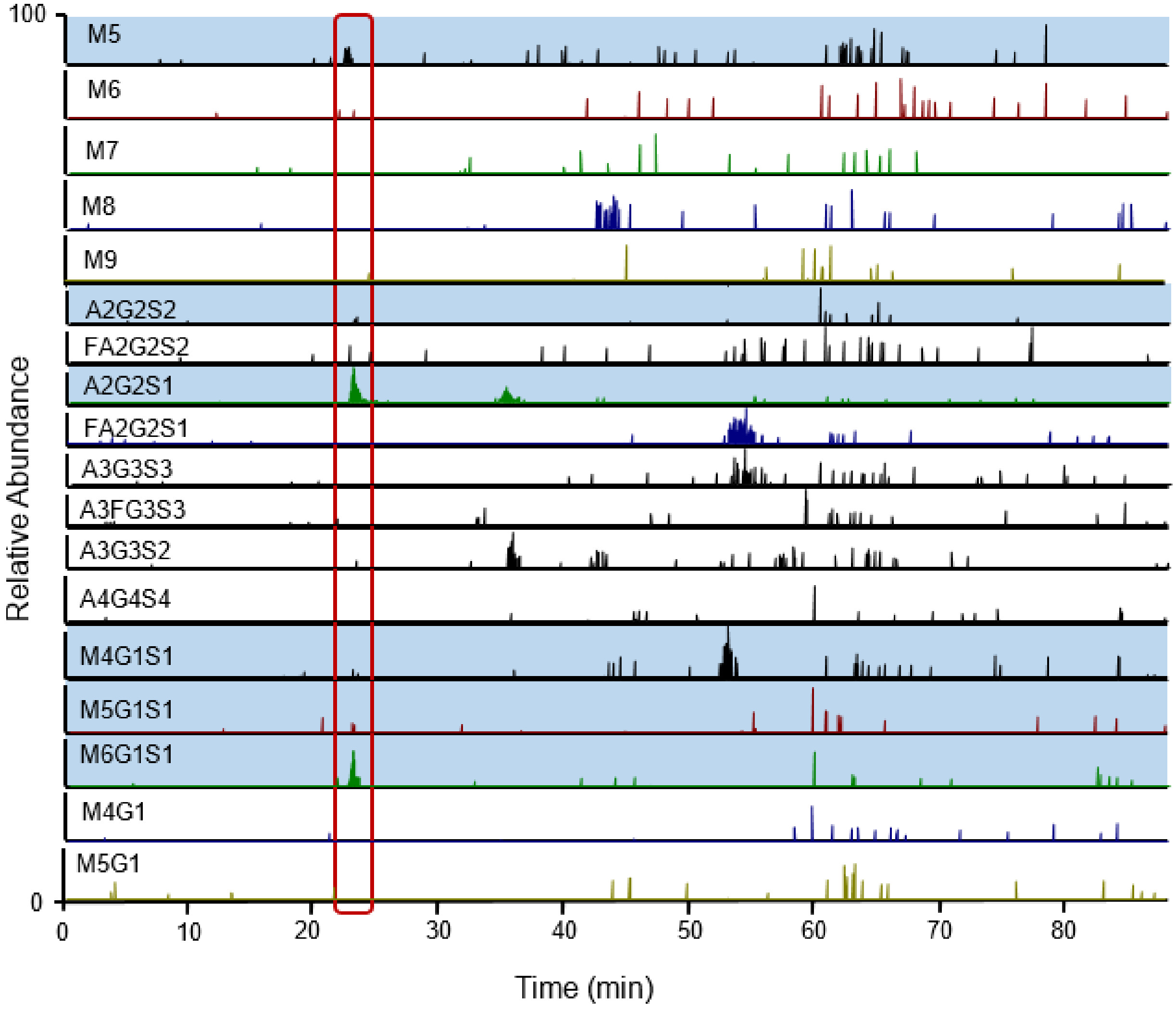


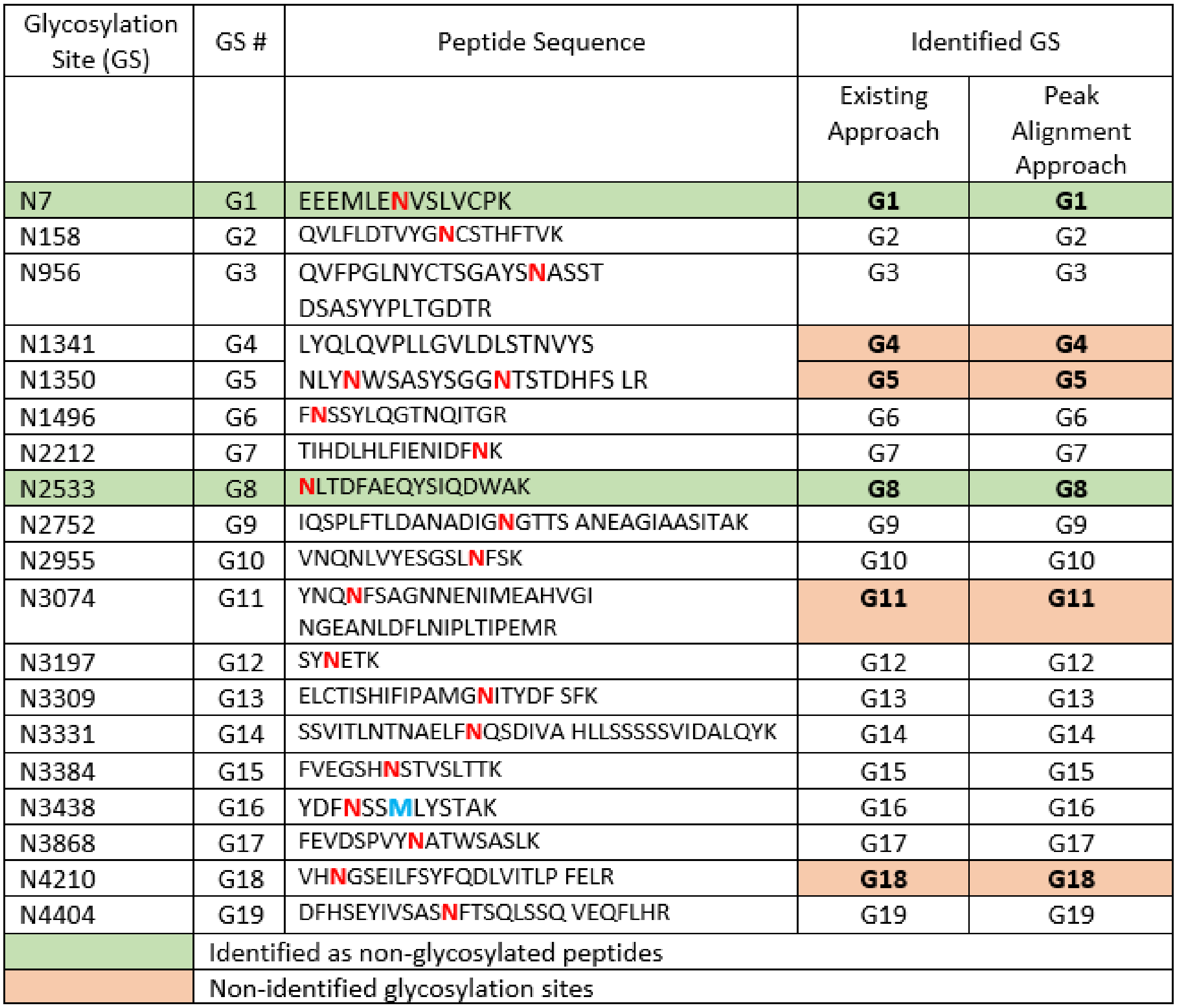
Figure 7. Representative data showing the application of the peak alignment approach to identify the N1496 (G6) glycosylation site of apoB100. The XIC's of doubly charged and triply charged glycoforms obtained for the F**N**[1496]SSYLQGTNQITGR glycopeptide are shown. Co-elution of six different glycoforms, containing high-mannose, complex and hybrid-type glycans, highlighted in blue, were observed in the time range of (21.00 – 24.00) mins. The *N*-linked glycosylation site identification was verified and confirmed by full MS and CID data. For details on the identified glycoforms, see Table 2 of the Appendix A.

The extracted *m/z*'s of the G6 glycosylation site included doubly charged high-mannose and hybrid glycoforms, and both doubly and triply charged complex glycoforms. For the G6 glycosylation site, a clear co-elution pattern was observed within the time range of (21.00 – 24.00) mins, and the co-eluted glycoforms represent glycans of three *N*-linked glycan groups. In other words, the G6 site of apoB100 was found to be occupied by high-mannose (M5), complex (A2G2S2 and A2G2S1) and hybrid (M4G1S1, M5G1S1 and M6G1S1) glycans. To find the additional glycoforms attached to the G6 glycosylation site, the full MS data corresponding to the G6 glycosylation site was submitted to Glycomod. The software assisted in the identification of two other candidate glycoforms: monoantennary complex [Hex]4[HexNAc]3[NeuAc]1 and biantennary complex [Hex]4[HexNAc]4[NeuAc]1. These glycoform identifications were then confirmed with the available CID data, see Appendix A, Table 2.

2.3.7 ApoB100 Summary

The overall findings for ApoB100 are in Table 2, which includes a glycosylation site identification summary, and Appendix A, Table 2 includes all the identified glycoforms of apoB100.

Table 2. Apolipoprotein B100 glycosylation site identification summary.



Glycosylation Site (GS)	GS #	Peptide Sequence	Identified GS	
			Existing Approach	Peak Alignment Approach
N7	G1	EEEMLE**N**VSLVCPK	**G1**	**G1**
N158	G2	QVLFLDTVYG**N**CSTHFTVK	G2	G2
N956	G3	QVFPGLNYCTSGAYS**N**ASST DSASYYPLTGDTR	G3	G3
N1341	G4	LYQLQVPLLGVLDLSTNVYS NLY**N**WSASYSGG**N**TSTDHFS LR	**G4**	**G4**
N1350	G5		**G5**	**G5**
N1496	G6	F**N**SSYLQGTNQITGR	G6	G6
N2212	G7	TIHDLHLFIENIDF**N**K	G7	G7
N2533	G8	**N**LTDFAEQYSIQDWAK	**G8**	**G8**
N2752	G9	IQSPLFTLDANADIG**N**GTTS ANEAGIAASITAK	G9	G9
N2955	G10	VNQNLVYESGSL**N**FSK	G10	G10
N3074	G11	YNQ**N**FSAGNNENIMEAHVGI NGEANLDFLNIPLTIPEMR	**G11**	**G11**
N3197	G12	SY**N**ETK	G12	G12
N3309	G13	ELCTISHIFIPAMG**N**ITYDF SFK	G13	G13
N3331	G14	SSVITLNTNAELF**N**QSDIVA HLLSSSSSVIDALQYK	G14	G14
N3384	G15	FVEGSH**N**STVSLTTK	G15	G15
N3438	G16	YDF**N**SS**M**LYSTAK	G16	G16
N3868	G17	FEVDSPVY**N**ATWSASLK	G17	G17
N4210	G18	VH**N**GSEILFSYFQDLVITLP FELR	**G18**	**G18**
N4404	G19	DFHSEYIVSAS**N**FTSQLSSQ VEQFLHR	G19	G19
🟩	Identified as non-glycosylated peptides			
🟧	Non-identified glycosylation sites			

In total, thirteen of the nineteen sites were detectable as glycopeptides using the approach described herein. Two of the sites, G1 and G8, were not identifiable as glycosylated using the targeted method, and a quick check for the nonglycosylated forms of the peptide indicated that, indeed, these sites are nonglycosylated. The fact that the method did not identify glyocforms for these sites is evidence that it is not susceptible to false positive identification. For the sites that were identified, a total of 43 unique glycoforms on 13 glycosylation sites were characterized.

Many of the sites had remarkably homogenous glycosylation, with only two glycoforms present. This finding is consistent with prior studies; only seven of 19 tryptic glycopeptides were shown previously to contain more than two glycoforms.[29] In hindsight, this finding demonstrates the power of the targeted approach used herein. Glycosylation sites with only two glycoforms present would be expected to be very difficult to find using an approach centered on searching for only 18 glycoforms. Yet, by choosing the forms that abundantly appear in glycopeptides from human serum, the approach was remarkably successful. However, not all sites were identified. Among the four unidentified glycosylation sites: G4, G5, G11 are not routinely observed upon trypsin digestion of apoB100,[29] so the lack of data on these sites was not surprising. The G18 glycosylation site could be detected previously, but it was not identified in this study. To investigate this anomaly, the entire protein was deglycosylated, and the deglycosylated peptide was searched for in a separate LC-MS experiment. It was not detectable, even as a deglycosylated peptide. As this particular peptide is the most hydrophobic in the study – it eluted at 100 minutes in a prior analysis – perhaps the chromatographic conditions in this study were not optimal for its detection. Finally, the method described herein performed identically to the comparator method (in Figure 1A), where two samples need to be prepared and two LC-MS files are collected and analyzed. While the comparator method required twice the protein, twice the sample prep, twice the data, and more analysis time, it did not result in any more identifications. Therefore, the new method is preferable, due to its simplicity.

2.4 Conclusion

We developed a rapid and simple approach for the *N*-linked glycosylation site mapping of glycoproteins within a single experiment. The novel approach, which relies on high-resolution MS data and chromatographic retention time, effectively identifies *N*-linked glycosylation sites by tracking co-eluting glycoforms corresponding to a particular glycosylation site in a reverse phase column. We successfully applied the workflow to profile the *N*-linked glycosylation sites of a heavily glycosylated human plasma glycoprotein; it was useful for efficiently mapping many of the glycosylation sites. The results showed the method's utility in rapidly analyzing a complex glycoprotein using a single LC-MS experiment and limited analysis time.

2.5 Acknowledgements

This work was supported by NIH Grant R01GM103547 to Heather Desaire.

2.6 References

1. Dwek, R. A., *Chemical reviews* **1996,** *96* (2), 683-720.
2. Furukawa, K.; Ohkawa, Y.; Yamauchi, Y.; Hamamura, K.; Ohmi, Y.; Furukawa, K., *Journal of biochemistry* **2012,** *151* (6), 573-8.
3. Kajihara, Y.; Tanabe, Y.; Sasaoka, S.; Okamoto, R., *Chemistry (Weinheim an der Bergstrasse, Germany)* **2012,** *18* (19), 5944-53.
4. Taniguchi, N.; Korekane, H., *BMB reports* **2011,** *44* (12), 772-81.
5. Cheon, Y. P.; Kim, C. H., *Clinical and experimental reproductive medicine* **2015,** *42* (3), 77-85.
6. Zabczynska, M.; Pochec, E., *Postepy biochemii* **2015,** *61* (2), 129-37.
7. Blomme, B.; Van Steenkiste, C.; Callewaert, N.; Van Vlierberghe, H., *Journal of hepatology* **2009,** *50* (3), 592-603.
8. Christiansen, M. N.; Chik, J.; Lee, L.; Anugraham, M.; Abrahams, J. L.; Packer, N. H., *Proteomics* **2014,** *14* (4-5), 525-46.
9. Saldova, R.; Reuben, J. M.; Abd Hamid, U. M.; Rudd, P. M.; Cristofanilli, M., *Annals of oncology : official journal of the European Society for Medical Oncology* **2011,** *22* (5), 1113-9.
10. Freeze, H. H., *Glycobiology* **2001,** *11* (12), 129r-143r.
11. Dube, D. H.; Bertozzi, C. R., *Nature reviews. Drug discovery* **2005,** *4* (6), 477-88.
12. Gornik, O.; Lauc, G., *Disease markers* **2008,** *25* (4-5), 267-78.

13. Abd Hamid, U. M.; Royle, L.; Saldova, R.; Radcliffe, C. M.; Harvey, D. J.; Storr, S. J.; Pardo, M.; Antrobus, R.; Chapman, C. J.; Zitzmann, N.; Robertson, J. F.; Dwek, R. A.; Rudd, P. M., *Glycobiology* **2008,** *18* (12), 1105-18.

14. Bailey, U. M.; Jamaluddin, M. F.; Schulz, B. L., *Journal of proteome research* **2012,** *11* (11), 5376-83.

15. Dalpathado, D. S.; Desaire, H., *The Analyst* **2008,** *133* (6), 731-8.

16. Go, E. P.; Hewawasam, G.; Liao, H. X.; Chen, H.; Ping, L. H.; Anderson, J. A.; Hua, D. C.; Haynes, B. F.; Desaire, H., *Journal of virology* **2011,** *85* (16), 8270-84.

17. Wuhrer, M., *Glycoconjugate J.* **2013,** *30* (1), 11-22.

18. Leymarie, N.; Zaia, J., *Anal Chem* **2012,** *84* (7), 3040-8.

19. Zhu, Z.; Desaire, H., *Annual Review of Analytical Chemistry* **2015,** *8*, 463-483.

20. Bern, M.; Kil, Y. J.; Becker, C., *Current protocols in bioinformatics* **2012,** *Chapter 13*, Unit13.20.

21. Wu, S. W.; Pu, T. H.; Viner, R.; Khoo, K. H., *Anal Chem* **2014,** *86* (11), 5478-86.

22. Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H., *Anal Chem* **2012,** *84* (11), 4821-9.

23. Lynn, K. S.; Chen, C. C.; Lih, T. M.; Cheng, C. W.; Su, W. C.; Chang, C. H.; Cheng, C. Y.; Hsu, W. L.; Chen, Y. J.; Sung, T. Y., *Anal Chem* **2015,** *87* (4), 2466-73.

24. Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N., *Journal of proteome research* **2013,** *12* (8), 3652-66.

25. Wang, B.; Tsybovsky, Y.; Palczewski, K.; Chance, M. R., *Journal of the American Society for Mass Spectrometry* **2014,** *25* (5), 729-41.

26. Go, E. P.; Ding, H.; Zhang, S.; Ringe, R. P.; Nicely, N.; Hua, D.; Steinbock, R. T.; Golabek, M.; Alin, J.; Alam, S. M.; Cupo, A.; Haynes, B. F.; Kappes, J. C.; Moore, J. P.; Sodroski, J. G.; Desaire, H., *Journal of virology* **2017,** *91* (9).

27. Behrens, A. J.; Vasiljevic, S.; Pritchard, L. K.; Harvey, D. J.; Andev, R. S.; Krumm, S. A.; Struwe, W. B.; Cupo, A.; Kumar, A.; Zitzmann, N.; Seabright, G. E.; Kramer, H. B.; Spencer, D. I.; Royle, L.; Lee, J. H.; Klasse, P. J.; Burton, D. R.; Wilson, I. A.; Ward, A. B.; Sanders, R. W.; Moore, J. P.; Doores, K. J.; Crispin, M., *Cell reports* **2016,** *14* (11), 2695-706.

28. Zhao, J.; Liu, Y. H.; Reichert, P.; Pflanz, S.; Pramanik, B., *Journal of mass spectrometry : JMS* **2010,** *45* (12), 1416-25.

29. Harazono, A.; Kawasaki, N.; Kawanishi, T.; Hayakawa, T., *Glycobiology* **2005,** *15* (5), 447-62.

30. Hwang, H.; Lee, J. Y.; Lee, H. K.; Park, G. W.; Jeong, H. K.; Moon, M. H.; Kim, J. Y.; Yoo, J. S., *Analytical and bioanalytical chemistry* **2014,** *406* (30), 7999-8011.

31. Song, T.; Aldredge, D.; Lebrilla, C. B., *Anal Chem* **2015,** *87* (15), 7754-62.

32. Thaysen-Andersen, M.; Mysling, S.; Hojrup, P., *Anal Chem* **2009,** *81* (10), 3933-43.

2.7 Appendix A

Table 1. *N*-linked glycan compositions reported for human plasma proteins tabulated in site-specific manner

Plasma Protein			Alpha-1-acid glycoprotein					Alpha-1-antitrypsin			Alpha-1B-glycoprotein				Alpha-2HS-glycoprotein		Alpha-2-macroglobulin								Antithrombin-III			
GS			N33	N56	N72	N93	N103	N70	N107	N271	N44	N179	N363	N371	N156	N176	N55	N70	N247	N396	N410	N869	N991	N1424	N128	N167	N187	N224
High Mannose		Man5																				🟩						
		Man6																				🟩						
		Man7																				🟩						
		Man8																										
		Man9																										
Complex	Bi-antennary	A2G2S2		🟩				🟩	🟩	🟩					🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩
		FA2G2S2						🟩	🟩	🟩					🟩	🟩	🟩							🟩			🟩	
		FA2BG2S2																										
		A2FG2S2																										
		FA2FG2S2																										
		A2G2S1											🟩												🟩	🟩	🟩	🟩
		FA2G2S1																										
		A2FG2S1																										
		FA2BG2S1																										
		A2BG2S1																										
		A2G2FS1																										
		A2																										
	Tri-antennary	A3G3S3	🟩	🟩	🟩	🟩	🟩		🟩																	🟩	🟩	
		A3FG3S3	🟩	🟩		🟩	🟩		🟩																			
		FA3FG3S3																										
		A3G3S2	🟩	🟩																								🟩
		A3FG3S2																										
		A3																										
		A3G3S1																										
		A3FG3S1																										
		FA3G3S2																										
		FA2FG2S2																										
	Tetra-antennary	A4G4S4			🟩	🟩	🟩																					
		A4FG4S4			🟩	🟩																						
		A4G4S3			🟩	🟩	🟩																					
		A4FG4S3			🟩																							
		A4FG3S3				🟩																						
		A4G4S2			🟩	🟩	🟩																					
		A4G4S1																										
Hybrid	Hybrid	M4G1S1																										
		M5G1S1																										
		M6G1S1																										
		M4G1																										
		M5G1																										

Plasma Protein			Apolipoprotein D		Apolipoprotein F			Beta-2-glycoprotein1				Ceruloplasmin							Fibrinogen	Fibrinogen alpha-chain		Fibrinogen beta-chain	Fibrinogen Gamma-chain		Haptoglobin			
GS			N65	N98	N118	N139	N267	N162	N183	N193	N253	N138	N227	N358	N397	N588	N762	N926		N453	N686	N394	N78	N334	N184	N207	N211	N241
High Mannose		Man5			🟩	🟩																						
		Man6			🟩	🟩																						
		Man7			🟩	🟩																						
		Man8			🟩	🟩																						
		Man9			🟩	🟩																						
Complex	Bi-antennary	A2G2S2	🟩	🟩				🟩		🟩		🟩		🟩	🟩		🟩					🟩	🟩	🟩	🟩	🟩	🟩	🟩
		FA2G2S2										🟩		🟩	🟩		🟩											
		FA2BG2S2																										
		A2FG2S2		🟩																					🟩	🟩		🟩
		FA2FG2S2																										
		A2G2S1						🟩		🟩												🟩	🟩	🟩	🟩	🟩		🟩
		FA2G2S1		🟩																								
		A2FG2S1																							🟩	🟩		🟩
		FA2BG2S1																										
		A2BG2S1																										
		A2G2FS1		🟩																								
		A2							🟩		🟩																	
	Tri-antennary	A3G3S3	🟩					🟩		🟩		🟩		🟩	🟩		🟩								🟩		🟩	🟩
		A3FG3S3		🟩								🟩			🟩		🟩								🟩		🟩	
		FA3FG3S3										🟩					🟩											
		A3G3S2	🟩					🟩		🟩															🟩		🟩	🟩
		A3FG3S2		🟩																								🟩
		A3							🟩		🟩																	
		A3G3S1																							🟩	🟩		🟩
		A3FG3S1																								🟩		
		FA3G3S2																										
		FA2FG2S2																										
	Tetra-antennary	A4G4S4	🟩														🟩											
		A4FG4S4															🟩											
		A4G4S3																										
		A4FG4S3		🟩																								
		A4FG3S3																										
		A4G4S2																								🟩		
		A4G4S1																								🟩		🟩
Hybrid	Hybrid	M4G1S1																										
		M5G1S1																										
		M6G1S1																										
		M4G1																										
		M5G1																										

Plasma Protein			Hemopexin					Serotransferrin			Vitronectin			Zinc-alpha-2-glycoprotein				Immunoglobulin A	IgA1		Immunoglobulin D			Immunoglobulin E							Immunoglobulin M					No. of times each glycan appeared	No. of different proteins
GS			N64	N187	N240	N246	N453	N432	N491	N630	N86	N169	N242	N109	N112	N128	N259		N144	N340	N225	N316	N367	N21	N49	N99	N146	N252	N264	N275	N46	N209	N272	N279	N439		
High Mannose		Man5																												🟩				🟩		5	4
		Man6																												🟩				🟩	🟩	6	4
		Man7																												🟩					🟩	5	4
		Man8																			🟩									🟩					🟩	5	4
		Man9																			🟩									🟩						4	3
Complex	Bi-antennary	A2G2S2	🟩	🟩			🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩		🟩																	45	15
		FA2G2S2	🟩	🟩			🟩			🟩			🟩							🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩								14	10
		FA2BG2S2																		🟩		🟩	🟩	🟩	🟩		🟩	🟩								7	3
		A2FG2S2						🟩		🟩																										6	3
		FA2FG2S2								🟩																										1	1
		A2G2S1						🟩		🟩		🟩	🟩				🟩		🟩			🟩	🟩													21	10
		FA2G2S1																		🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩			12	5
		A2FG2S1																																		3	1
		FA2BG2S1																						🟩	🟩			🟩			🟩	🟩	🟩			6	2
		A2BG2S1																	🟩																	1	1
		A2G2FS1																																		1	1
		A2																																		2	1
	Tri-antennary	A3G3S3									🟩		🟩																							20	8
		A3FG3S3									🟩		🟩																							13	6
		FA3FG3S3																																		2	1
		A3G3S2						🟩		🟩																										11	6
		A3FG3S2								🟩																										3	3
		A3																																		2	1
		A3G3S1																																		3	1
		A3FG3S1																																		1	1
		FA3G3S2								🟩																										1	1
		FA2FG2S2								🟩																										1	1
	Tetra-antennary	A4G4S4																																		5	3
		A4FG4S4																																		3	2
		A4G4S3																																		3	1
		A4FG4S3																																		2	2
		A4FG3S3																																		1	1
		A4G4S2																																		4	2
		A4G4S1																																		2	1
Hybrid	Hybrid	M4G1S1										🟩																								1	1
		M5G1S1										🟩																								1	1
		M6G1S1																																		0	0
		M4G1										🟩																								1	1
		M5G1										🟩																								1	1

Table 2. *N*-linked glycopeptides of Apolipoprotein B100 verified with the Full MS and confirmed with the CID data

N-linked Glycosylation Site	ID	Peptide Mass/ Da	CS	Theoretical *m/z*	Observed *m/z*	Mass Error /ppm	Glycan Notation	Approx. Retention Time/min
EEEMLEN^{7}VSLVCPK	G1	1675.7797	2+	838.8972	838.9012	4.8	NG	**26.76**
			3+	559.6005	559.6029	4.3	NG	
QVLFLDTV YGN^{158} CSTHFTVK	G2	2228.0936	3+	1149.1795	1149.1850	4.8	M5	**32.34**
			3+	1203.1971	1203.2024	4.4	M6	
			3+	1257.2147	1257.2188	3.3	M7	
			3+	1311.2323	1311.2352	2.2	M8	
			3+	1365.2499	1365.2533	2.5	M9	
QVFPGLNYCTSGAYSN^{956}ASSTDSASYYPLTGDT	G3	3549.5630	3+	1919.1191	1919.1212	1.1	A2G2S2	**32.62**
			3+	1822.0873	1822.0907	1.9	A2G2S1	
LYQLQVPLLGVLDLSTNVYSNLYN^{1341}WSASYSGG NTSTDHFSLR	G4	4692.3136					ND	
LYQLQVPLLGVLDLSTNVYSNLYNWSASYSGG N^{1350}TS TDHFSLR	G5	4692.3136					ND	
FN^{1496}SSYLQGTNQITGR	G6	1684.8169	2+	1451.6272	1451.6323	3.5	M5	**22.00**
			2+	1945.8020	1945.8002	0.9	A2G2S2	
			2+	1800.2543	1800.2543	1.3	A2G2S1	
			2+	1698.7146	1698.7104	2.5	M4G1S1	
			2+	1779.7410	1779.7676	15.0	M5G1S1	
			2+	1860.7674	1860.7956	15.2	M6G1S1	
			2+	1719.2279	1719.2325	2.7	A2G1S1	
			2+	1617.6882	1617.6925	2.7	A1G1S1	
TIHDLHLFIENIDFN^{2212}K	G7	1968.0105	3+	1391.9350	1391.9384	2.4	A2G2S2	**35.44**
			3+	1294.9032	1294.9080	3.7	A2G2S1	
N^{2533}LTDFAEQYSIQDWAK	G8	1927.8952	2+	964.9549	964.9586	3.8	NG	**36.42**
			3+	643.6390	643.6407	2.6	NG	
IQSPLFTLDANADIGN^{2752}GTTS ANEAGIAASITAK	G9	3231.6258	3+	1813.1401	1813.1487	4.7	A2G2S2	**39.80**
			3+	1716.1083	1716.1082	0.0	A2G2S1	
VNQNLVYESGSLN^{2955}FSK	G10	1797.8897	3+	1335.2280	1335.2318	2.8	A2G2S2	**24.81**
			3+	1238.1962	1238.2004	3.4	A2G2S1	
YNQN^{3074}FSAGNNENIMEAHVGINGEANLDFLNIPL TIPEMR	G11	4359.0688					ND	
SYN^{3197}ETK	G12	740.3341	2+	1473.5606	1473.5582	1.6	A2G2S2	**3.10**
			2+	1328.0129	1328.0128	0.1	A2G2S1	
ELCTISHIFIPAMGN^{3309}ITYDF SFK	G13	2703.3076	3+	1307.5841	1307.5810	2.4	M5	**38.42**
SSVITLNTNAELFN^{3331}QSDIVA HLLSSSSSVIDALQYK	G14	3863.9792	3+	1910.5451	1910.5503	2.7	M9	**54.57**
FVEGSHN^{3384}STVSLTTK	G15	1605.7998	3+	941.7482	941.7502	2.1	M5	**18.89**
			3+	1271.1981	1271.2014	2.6	A2G2S2	
			3+	1174.1663	1174.1704	3.5	A2G2S1	
			3+	1106.4731	1106.4787	5.0	M4G1S1	
			3+	1160.4907	1160.4945	3.3	M5G1S1	
			3+	1009.4413	1009.4440	2.7	M4G1	
			3+	1052.4555	1052.4593	3.6	A1G1S1	
			3+	1120.1487	1120.1481	0.5	A2G2S1	
YDFN^{3438}SSMLYSTAK*	G16	1541.6759	2+	1874.2315	1874.2327	0.6	A2G2S2	**22.64**
Methionine Oxidized			2+	1728.6838	1728.6845	0.4	A2G2S1	
			2+	1789.1969	1789.2199	12.8	M6G1S1	
			2+	1647.6574	1647.6554	1.2	A2G1S1	
			2+	1546.1177	1546.1170	0.5	A1G1S1	
FEVDSPVYN^{3868}ATWSASLK	G17	1912.9207	3+	1373.5717	1373.5751	2.5	A2G2S2	**30.47**
			3+	1276.5399	1276.5427	2.2	A2G2S1	
			3+	1495.2824	1495.2911	5.8	A3G3S2	
VHN^{4210}GSEILFSYFQDLVITLP FELR	G18	2836.4799					ND	
DFHSEYIVSASN^{4404}FTSQLSSQVEQFLHR	G19	3155.4948	3+	1787.7631	1787.7722	5.1	A2G2S2	**50.98**
			3+	1690.7313	1690.7377	3.8	A2G2S1	

NG	Non-glycosylated peptides
ND	Non-identified glycosylation sites
🟨	Additional glycoforms identified with Glycomod tool

Chapter 3. A Clinically Viable Assay for Monitoring Uromodulin Glycosylation

Abstract

Uromodulin, known as the Tamm-Horsfall protein or THP; the most abundant protein excreted in human urine, is associated with the progression of kidney diseases. Therefore, changes in the glycosylation profile of this protein could serve as a potential biomarker for kidney health. The typical glycomics analysis approaches used to quantify uromodulin glycosylation involve time-consuming and tedious glycoprotein isolation and labeling steps, which limit their utility in clinical glycomics assays. Herein we introduce a radically simplified sample preparation with direct ESI-MS analysis, enabling the quantitation of *N*-linked glycans that originate from uromodulin. The method omits any glycan labeling steps, but includes steps to reduce the salt content of the samples, to reduce the ion suppression. The method is effective for quantifying subtle glycosylation differences of uromodulin samples derived from different biological states. Furthermore, it provided highly reproducible quantitation data for within-group samples, which allow different samples from the same biological state to be classified together using PCA.

3.1 Introduction

Protein glycosylation, where glycans are covalently attached to the proteins through the side chains of certain amino acid residues, is one of the most abundant post translational modifications (PTMs) found in nature. Like other PTMs found on proteins, this modification introduces huge diversity not only to the protein structure, but also to its' functions.[1] However, protein glycosylation is highly sensitive to the changes in the cellular environment; yielding aberrantly glycosylated proteins during the progression of many diseases, such as cancers,[2-4] kidney diseases,[5-6] arthritis,[7] and Parkinson's disease.[8] Therefore, the relative abundances of these altered glycans often represent changes in biological states, such as healthy versus disease.[6-7] Hence, these altered glycans derived from complex biological fluids or a specific protein provides unique opportunities for disease diagnosis and prognosis.

One important example of a protein whose glycosylation could serve as a biomarker is uromodulin. Uromodulin, also known as the Tamm-Horsfall protein or THP, the most abundant glycoprotein excreted in human urine with a daily excretion rate of 50 – 100 mg,[5, 9-10] found to play important roles in preventing kidney stone formation[5] and urinary tract infections.[11] Uromodulin is 94 kDa in size and glycans represent approximately 25 - 30% of its weight. This glycoprotein contains eight potential *N*-linked glycosylation sites, of which 7 are reported to be glycosylated; these sites are mainly occupied by various complex-type di-, tri-, and tetra-antennary glycans, in addition to the minute level of high mannose-type glycans.[10, 12] One unique feature of uromodulin glycosylation profile is that the acidic nature of the many of the reported glycans; these glycans can contain sialic acids and/or sulfate substituents, such as 3-*O*-sulfated galactose (Gal3*S*) and/or 4-*O*-sulfated *N*-acetylgalactosamine (GalNAc4*S*).[9, 13-14] Analysis of these different glycans of uromodulin is important because of their significance in distinguishing samples of various biological states; for examples, reduced levels of overall glycosylation and sialylation of uromodulin glycoprotein was reported in patients with interstitial cystitis[10] and kidney stones.[5] Therefore, development of efficient methods for the sensitive detection of uromodulin glycans is important in clinical studies, as they can serve as critical biomarkers for various diseases, while allowing discriminating of samples of different health states.

While analysis of uromodulin glycosylation has the potential to improve diagnosis and treatment of a variety of kidney-related conditions, a simple, and accurate assay that would be clinically viable is not currently available. The purification of the protein from urine is currently done using complex sample preparation procedures; using either diatomaceous earth,[5, 15] or salt precipitation.[10, 12] Furthermore, once the glycans are purified and released, general application of existing glycomics assays introduces many more additional steps; these steps typically include glycan labeling and post-sample clean-up steps, which are laborious and time-consuming. To resolve the challenges of laborious sample preparation methods described above, and to provide kidney researchers with the opportunity to readily assess uromodulin glycosylation changes for improving the diagnosis and prognosis of kidney diseases, we developed a clinically viable procedure for the analysis of uromodulin glycans.

Because uromodulin is, by far, the most abundant protein in urine,[5, 9-10] it is possible to develop a radically simplified procedure to generate highly enriched uromodulin samples without the need for purification from diatomaceous earth, which is the most common protocol. In the protocol described below, all proteins below 50 kD are removed using a molecular weight cut-off filter, removing many potential low molecular weight interferents. IgG is the next most abundant in urine after uromodulin,[16-20] but its concentration is still ~ 100X lower than that of uromodulin,[16] so its glycosylation in general would minimally impact this assay. To further reduce the minimal impact of IgG, the uromodulin analysis is exclusively conducted in the negative ion mode, which is optimal for uromodulin glycans but a poor choice if the goal is to detect IgG glycans, since IgG’s main glycoforms are neither sialylated nor sulfated.[21-23] Overall, this enrichment procedure and analysis in negative ion mode optimizes the balance between the need for samples that are highly enriched in uromodulin with the need for a radically simplified workflow that could be applied on large sample sets.

Aside from simplifying the protein enrichment step, the other aspect of the work described herein, which is necessary to advance researchers' ability to analyze uromodulin on large banks of clinical samples, is to address the (typically many) labeling steps that are generally thought to be necessary prior to a quantitative glycomics analysis. These steps usually involved reductive amination (a two-step reaction that generates hazardous waste), removal of additional derivatization reagents, and then analysis of the labeled glycans typically with LC-Fluorescence and MALDI-MS.[5, 10, 12, 24]

We hypothesized that since uromodulin is available in abundant quantities, the typical glycan labeling and enrichment steps would not be necessary. While the analyte could be analyzed at relatively high concentrations, because enough of it was available, the salts that were present in the analysis would have to be removed or minimized. We therefore developed an efficient strategy for salt removal that could be applied on large sample sets. The method was first developed using fetuin as a model glycoprotein to demonstrate the generality of the approach for analyzing for acidic glycoproteins, then it was applied to uromodulin. It proved to be highly reproducible over multiple sample preparations and multiple analyses. It was successfully applied to quantify *N*-linked glycans of the uromodulin standards, followed by quantitation of *N*-linked glycans of uromodulin, enriched from human urine samples of two different biological states. The method provided highly reproducible quantitation results over multiple samples of the same biological state resulting very tight within-group clustering with PCA, while showing its potential utility in classifying samples of patients with kidney diseases based on uromodulin-specific glycan biomarkers.

3.2 Experimental

3.2.1 Materials and Reagents

Bovine fetuin and human uromodulin standards were purchased from Sigma Aldrich (St. Louis, MO) and BioVendor (Asheville, NC), respectively. PNGase F (500, 000 units/mL) was from New England Biolabs (Ipswich, MA). Single donor human urine from a de-identified healthy female and a de-identified third trimester pregnant female were purchased from Innovative Research (Novi, MI). All the chemical reagents used for this study were of analytical grade or better.

3.2.2 PNGase F Enzyme Preparation

PNGase F (1 µL, 500 units) was diluted to 100 µL (for fetuin) or PNGase F (2 µL, 1000 units) was diluted to 200 µL (for uromodulin and urine samples) with NH_4HCO_3 buffer (10 mM, pH 7.5) and concentrated in a pre-rinsed 10 kD MWCO filter (14000 g × 15 min) to approximately 50 µL of final volume. Then, the concentrated enzyme solution was diluted by a factor of 10 with the buffer (10 mM NH_4HCO_3, pH 7.5) followed by another concentrating step (14000 g × 15 min) to obtain approximately 35 µL of final enzyme solution. Finally, the PNGase F concentrate was collected through reverse spin (1000 g × 2 min).

3.2.3 *N*-linked Glycan Release from Fetuin and Uromodulin Standard

Glycoprotein samples (50 µg) were obtained by transferring appropriate volumes of fetuin (5 mg/mL in 10 mM NH_4HCO_3 buffer) and uromodulin (2 mg/mL in water) glycoprotein stock solutions. Then, the final glycoprotein solution concentration was adjusted to 2 mg/mL by diluting with the buffer (for fetuin). The PNGase F enzyme solution (35 µL), prepared as described in above, was added to each glycoprotein solution and incubated for 24 hours at 37 $^{\circ}$C. After the incubation period, samples were diluted up to 500 µL with 50:50 MeOH: H_2O, transferred to pre-rinsed 10 kD MWCO filters, and centrifuged (14000 g × 15 min) to collect the filtrate with released *N*-linked glycans. The resultant filtrates were then concentrated in a centrivap vacuum concentrator, to yield a final volume of 15 µL, and stored at -20 °C. Prior to the direct ESI-MS analysis, *N*-linked glycans' concentrate solutions of fetuin and uromodulin were diluted approximately 70 and 17 times with 50:50 MeOH: H2O, respectively.

3.2.4 Enrichment of Uromodulin from Human Urine and *N*-linked Glycan Release

Uromodulin was enriched from the crude urine samples obtained from a healthy and a pregnant individual. Urine samples, which were stored at - 80 °C, were thawed to room temperature, vortexed for 10 seconds, and then aliquoted into 10 mL fractions. Three of the 10 mL aliquots of each urine sample (healthy and pregnant) were vortexed for 10 seconds, followed by transferring the vortexed urine samples into pre-rinsed Amicon 50 kD molecular weight cut-off filters (Millipore), separately. The samples were centrifuged at 3500 rpm for 15 mins (at 4 °C) to reduce the volume approximately down to 300 µL. Then the protein concentrate was washed by adding the buffer up to 15 mL and the samples were centrifuged at 3500 rpm for 15 mins. The washing step was repeated another two times to obtain the final concentrate (~ 300 µL) for each triplicate samples; the retentate was carefully transferred to a pre-rinsed 50 kD MWCO filter (0.5 mL); centrifuged at 7000 rpm for 30 mins to further reduce the sample volume to 50 µL. Then the final concentrate was collected by performing reverse spin at 1000 g for 2 min. After that, the *N*-linked glycan release was performed by adding salt-reduced PNGase F enzyme (1000 units), which was subjected to 100 times of dilution with the buffer, centrifugation, and second-buffer exchange step as described above. Thereafter, the samples were incubated, *N*-linked glycans were collected, samples were concentrated, and stored as described for the protein standards. Finally, the concentrated *N*-linked glycan solutions, which were generated from triplicate samples of a healthy and a pregnant urine, were diluted 17 times with 50:50 MeOH: H2O (v/v), prior to the direct ESI-MS analysis.

3.2.5 Direct ESI-MS Analysis of Released *N*-linked Glycans

Direct ESI-MS analysis of the released *N*-linked glycans was performed using an Orbitrap Fusion Tribrid mass spectrometer (Thermoscientific, San Jose, CA). The mass spectrometer was operated in negative ion mode with a sample injection flow rate of 5 µL/min. The heated-electrospray source was held at 2.4 kV while the ion transfer tube temperature, sheath and auxiliary gas flow rates were set at 300 ˚C, 10, and 8 Arb units, respectively. The full MS scans for the *m/z* range of (750 – 1600) were acquired in the Orbitrap with a resolution of 120 k (at *m/z* 200). The AGC target value for the full MS scan was 4×10^5, and the maximum injection time was 100 ms. For full MS data of fetuin and uromodulin standards, 30 scans were averaged, while 100 scans were averaged during the analysis of uromodulin *N*-linked glycans extracted from urine samples.

Both CID and HCD data for the *N*-linked glycans derived from fetuin and uromodulin standards, were generated; some additional MS/MS data were acquired by using uromodulin; extracted from normal and pregnant urine samples, to confirm the glycan compositions those were not observed in the uromodulin standard samples. For MS/MS, the isolation window for the precursor ions were set as 2 Da, activation time, and activation q_z were 10 ms and 0.25, respectively. The selected precursor ions were fragmented by applying appropriate normalized collision energies ranging between 30% - 35%. All the MS data were acquired by using Xcalibur software, version 4.2 (Thermoscientific, San Jose, CA).

3.2.6 Data Analysis

Appendix B, Table 1 represents a list of 40 potential uromodulin *N*-linked glycans tabulated from previous reports.[5, 9, 12-14, 25] The resultant full MS data were searched against the *N*-linked glycans listed in Table 1 of the Appendix B, for glycan assignment; these assignments were done by comparing the theoretical monoisotopic masses of listed *N*-linked glycans with their experimental *m/z*'s, and identified the *N*-linked glycans within 5 ppm mass error. These assignments were further confirmed by analyzing the resultant MS/MS data.

3.2.6.1 Quantifying the Glycans

Full MS scans; 30 and 100 scans for *N*-linked glycan samples derived from standard glycoproteins and urine samples, respectively, were averaged. The first four isotopic peaks' raw abundances of each *N*-linked glycan were summed over all the identified charge states and adducts (protonated forms and sodiated adducts). Then, the percent of each glycan, based on its peak intensity was calculated by dividing a particular glycan peak intensity by the total *N*-glycan peak intensity of the analyzed sample, multiplied by 100.[26]

3.2.6.2 Classification of Sample Groups

Principal Component Analysis (PCA) was conducted in R, version 3.5.1, using the package "factoextra". The data were centered and scaled prior to the PCA transformation.

3.3 Results and Discussion

3.3.1 Overview of the Label-free *N*-linked Glycan Quantitation Approach

Sample preparation is one critical step in the clinical biomarker discovery field that affects the final throughput of the method. Therefore, development of simple and efficient sample preparation strategies is necessary in quantitative glycomics analysis. Figure 1 shows a schematic diagram of a simple *N*-linked glycan preparation protocol that enables efficient release and direct ESI-MS quantitation of *N*-linked glycans without labeling.

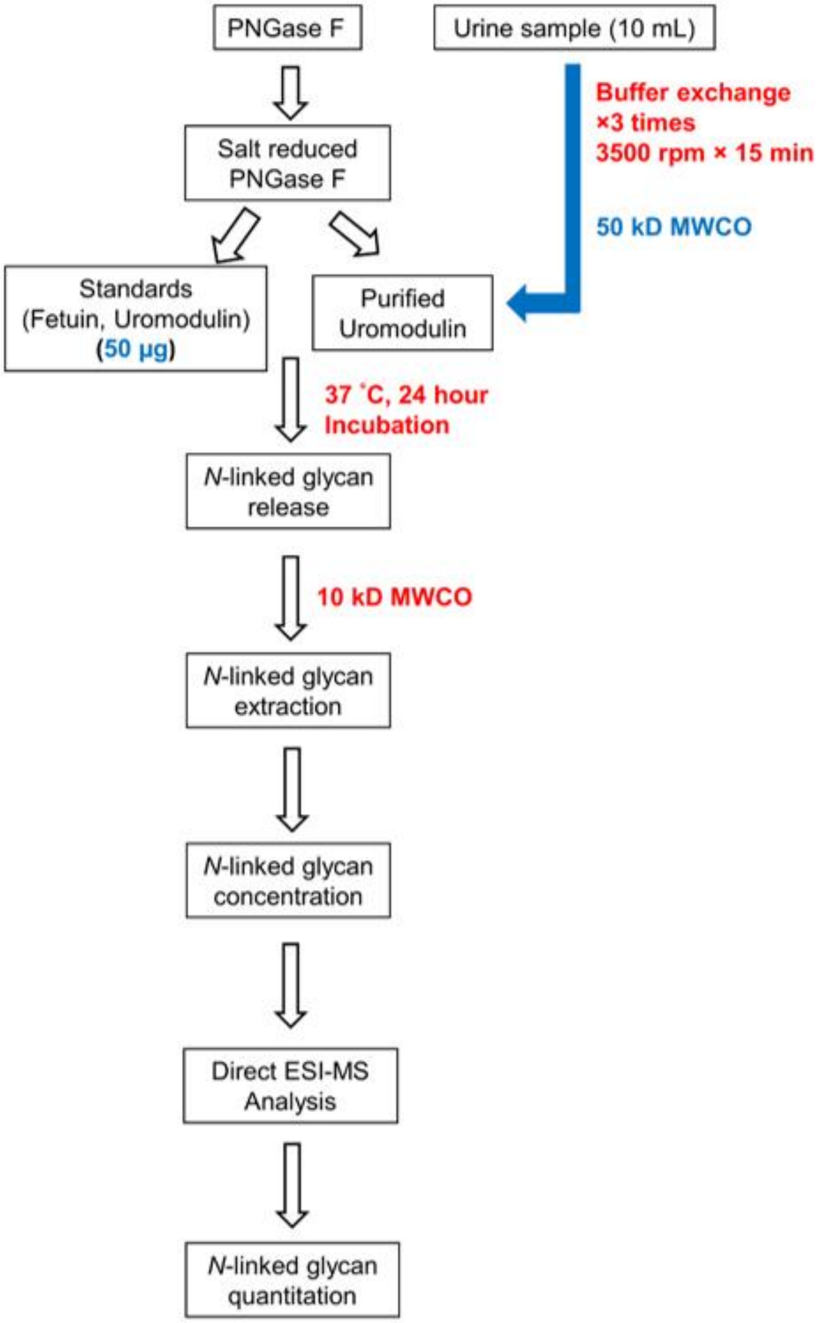


Figure 1. Experimental Workflow for *N*-linked glycan profiling of fetuin and uromodulin standard glycoproteins and uromodulin, extracted from urine samples.

In this protocol, as the first step, *N*-linked glycans are released from the glycoproteins of interest by incubating them with salt-reduced PNGase F enzyme. In direct ESI-MS analysis, the presence of salts can increase the ion suppression, reduce the stability of electrospray, and affect the sensitivity of the analysis.[27-29] Thus, reducing the amount of salt present in the samples being analyzed is important prior to the direct ESI-MS analysis. Both urine samples and PNGase F, used in this study contained salts, thus, these samples needed to be desalted prior to any other sample processing steps. Urine samples typically contain high salt concentration;[30] thus, they were desalted by following several washing steps as shown in Figure 1. Additionally, PNGase F, which contains 50 mM of NaCl, was also buffer exchanged several times to minimize the initial salt concentration by at least three orders of magnitude before adding to the glycoprotein solution. Then, after a 24 h incubation period, the glycans were directly extracted, concentrated and analyzed with direct ESI-MS in the negative ion mode without prior labeling.

This protocol differs from other standard glycomics approaches because it omits a labeling step. Glycan labeling prior to the MS analysis improves glycan's ionization efficiency;[1, 31-34] however, the labeling process and post-sample clean up steps are time-consuming and potentially introduce additional variability into the analysis. The developed protocol is more rapid, while allowing highly reproducible quantification of the relevant *N*-linked glycans derived from standard glycoproteins and uromodulin, enriched from human urine samples via direct ESI-MS in the negative ion mode.

The ultimate goal of this study was to develop a simple, label-free, direct ESI-MS approach to effectively profile *N*-linked glycosylation of uromodulin glycoprotein, which is mainly occupied by negatively charged glycans those ionize well in the negative ion mode. However, as this glycoprotein has a complex *N*-linked glycosylation profile, we performed the initial method development on a standard glycoprotein that has a simpler glycosylation profile than the uromodulin, yet, contains negatively charged glycans those ionize well in the negative ion mode. Therefore, we selected fetuin as the standard glycoprotein for the method development and optimization, and the generated results were used to test the method reproducibility and the instrument precision, as described below.

3.3.2 Reproducibility of the Method

Higher reproducibility of a method generally permits enhanced sensitivity towards differentiating minor changes across multiple samples with a higher confidence.[35] If the method is reproducible, small differences that are introduced during the sample preparation steps, such as PNGase F release, extraction of *N*-linked glycans, and dilution of concentrated *N*-linked glycan samples prior to the MS analysis, should not affect the final quantitation results. Therefore, to test the reproducibility of the quantitative sample preparation workflow, fetuin (50 µg) obtained from the same stock solution was subjected to *N*-linked glycan release protocol on three different days and analyzed by (-)ESI-MS under identical instrumental parameters, as described in the experimental section. Figure 2 represents a direct ESI-MS *N*-linked glycan profile derived from the fetuin *N*-linked glycan sample, in the negative ion mode; the resulted fetuin *N*-linked glycans were assigned across multiple charge states and multiple adducts (protonated and sodiated).

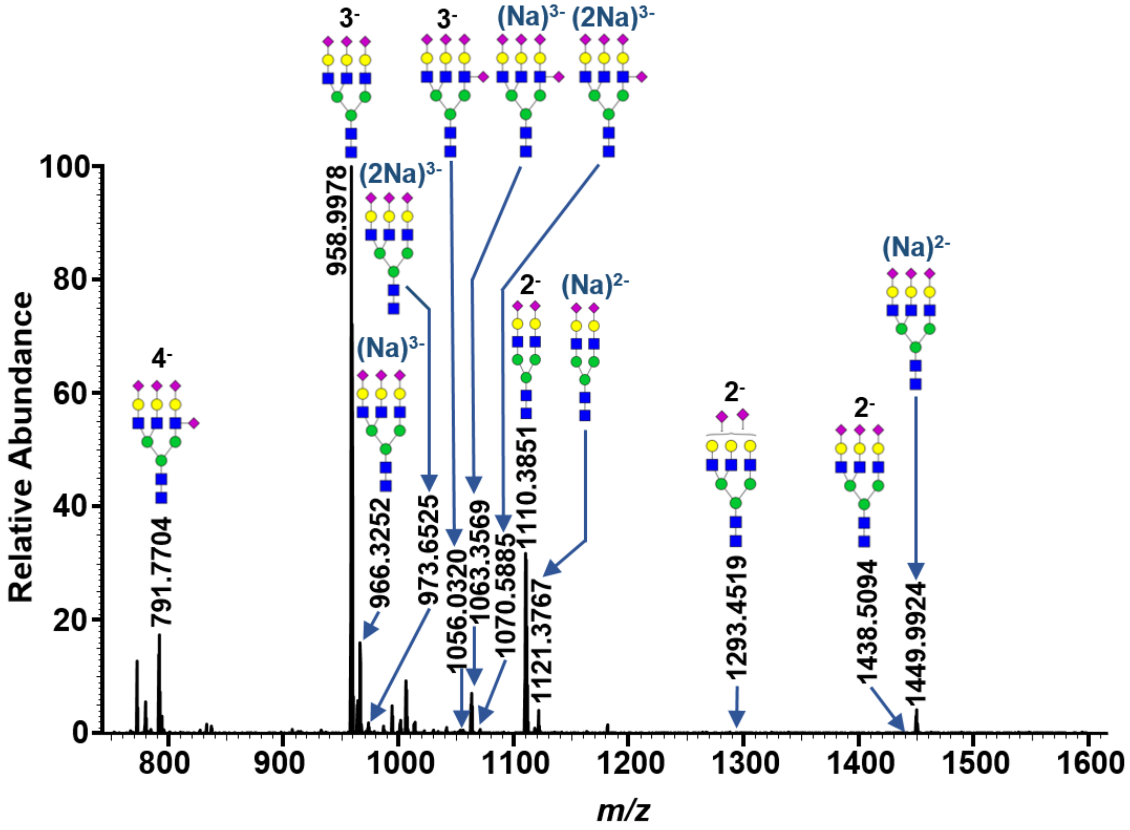


Figure 2. Negative ion mode direct ESI-MS spectrum of released label-free *N*-linked glycan profile of fetuin sample. The protonated and sodium adducts are indicated across multiple charge states. Monosaccharide symbol key: blue square (*N*-acetylglucosamine); green circle (mannose); yellow circle (galactose); pink diamond (*N*-acetylneuraminic acid).

After obtaining the *N*-linked glycan profile of fetuin, the relative glycan peak percent of the individual glycans were calculated as described in the data analysis section. See Figure 3A, which includes data for four different complex-type *N*-linked glycans of fetuin over three separate sample preparations. In every sample preparation, the observed relative glycan peak percentage differences across major *N*-linked glycans were subtle.

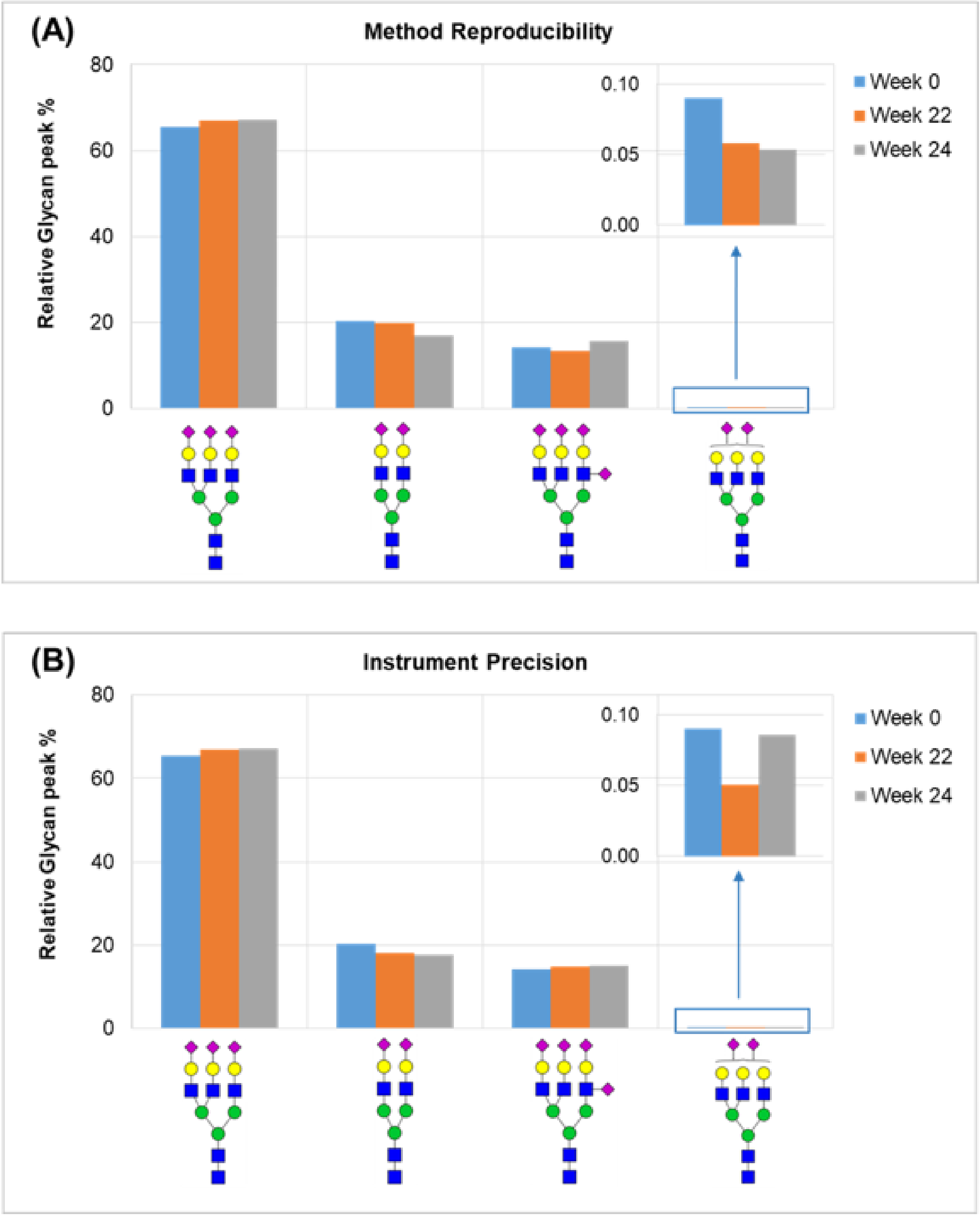


Figure 3. Method reproducibility **(A)** and instrument precision **(B)** calculated for four different fetuin *N*-linked glycans. The relative glycan peak percent for each *N*-linked glycan composition is plotted for three different sample preparations from the same stock **(A)** and for the same sample analyzed over week 0, week 22, and week 24 **(B)**. The *N*-linked glycans are rank ordered from the largest percentage to the smallest percentage. Less than 8% of CV values were recorded for 3 major fetuin *N*-linked glycans and about 24% of CV was calculated for the least abundant fetuin *N*-linked glycan in both **A** and **B**. The relative glycan peak percentages recorded for the least abundant fetuin *N*-linked glycan are zoomed in for both **(A)** and **(B)**.

Table 1A summarizes the raw abundances, mean relative glycan peak percentages, and coefficients of variation values calculated for four fetuin *N*-linked glycans: H6N5S3, H5N4S2, H6N5S4, and H6N5S2; the glycan abundances were 67%, 19%, 14%, and 0.067% for the four fetuin *N*-linked glycans, respectively. Additionally, the coefficient of variation values were 1.2%, 7.7%, 7.1%, and 24%, respectively. The results showed that the method is highly reproducible while showing less than 8% of coefficients of variation for all major fetuin *N*-linked glycans, except for the least abundant H6N5S2 glycan peak, which represented less than 1% of the sample.

Table 1. Raw abundances, mean relative glycan peak percentages and coefficients of variation values calculated for four fetuin *N*-linked glycans; over multiple sample preparations for method reproducibility **(A)** and over multiple analysis of a single fetuin sample for instrument precision **(B)**.

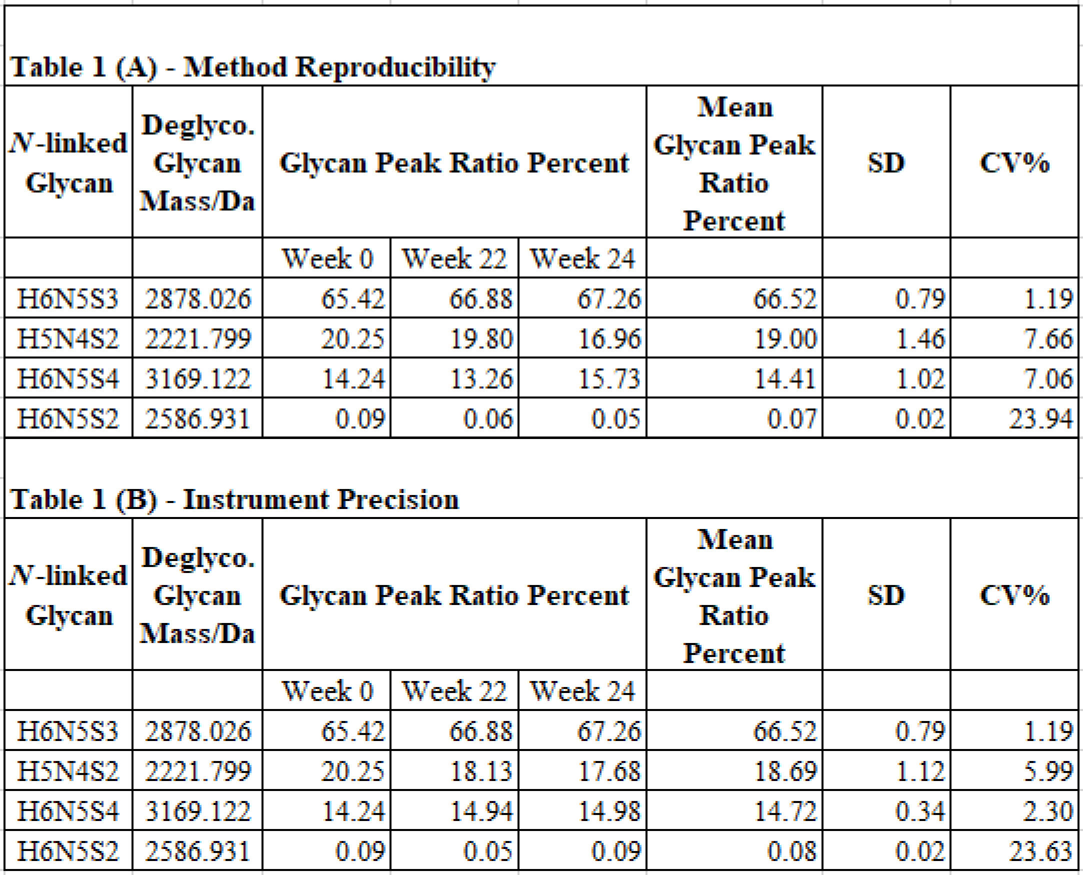


Table 1 (A) - Method Reproducibility

N-linked Glycan	Deglyco. Glycan Mass/Da	Glycan Peak Ratio Percent			Mean Glycan Peak Ratio Percent	SD	CV%
		Week 0	Week 22	Week 24			
H6N5S3	2878.026	65.42	66.88	67.26	66.52	0.79	1.19
H5N4S2	2221.799	20.25	19.80	16.96	19.00	1.46	7.66
H6N5S4	3169.122	14.24	13.26	15.73	14.41	1.02	7.06
H6N5S2	2586.931	0.09	0.06	0.05	0.07	0.02	23.94

Table 1 (B) - Instrument Precision

N-linked Glycan	Deglyco. Glycan Mass/Da	Glycan Peak Ratio Percent			Mean Glycan Peak Ratio Percent	SD	CV%
		Week 0	Week 22	Week 24			
H6N5S3	2878.026	65.42	66.88	67.26	66.52	0.79	1.19
H5N4S2	2221.799	20.25	18.13	17.68	18.69	1.12	5.99
H6N5S4	3169.122	14.24	14.94	14.98	14.72	0.34	2.30
H6N5S2	2586.931	0.09	0.05	0.09	0.08	0.02	23.63

3.3.3 Instrument Precision

Label-free quantitative assays performed with mass spectrometers can be subjected to reproducibility issues over lengthy time periods, as a result of slight changes occurring in the instrument conditions.[26] Therefore, the instrument precision over the time period of the analysis was tested by analyzing a fetuin sample on three different days: at week 0, week 22, and week 24. After the first analysis performed on week 0, the released *N*-linked glycans from the sample were stored at -20 °C and re-analyzed in week 22 and week 24 under identical ESI-MS conditions.

Table 1B shows the recorded raw abundances, mean relative glycan peak percentages, and coefficients of variation values calculated for four fetuin *N*-linked glycans. The rank order recorded for the fetuin *N*-linked glycans were consistent over the analysis at different time points, and the coefficients of variation of relative glycan peak percentages calculated across all the *N*-linked glycans were lower than 6%, except for the least abundant glycan peak: H6N5S2, which showed about 24% of coefficient of variation. Figure 3B illustrates the relative glycan peak percentages recorded for four fetuin *N*-linked glycans across three time points, and these data clearly show that the instrument performance remained unchanged during the time period of the study.

Based on the initial quantitative data obtained with the fetuin *N*-linked glycans, the method showed to be highly reproducible over multiple sample preparations and under the MS conditions used for the study. Therefore, we tested the applicability of the developed label-free direct ESI-MS method towards efficient quantitation of *N*-linked glycans derived from a more complex, glycoprotein, uromodulin.

3.3.4 Quantification of Human Uromodulin *N*-linked Glycans

The *N*-linked glycans of human uromodulin were released and extracted from 50 µg of a purchased glycoprotein standard; the released glycans were concentrated, diluted, and analyzed directly by ESI-MS. This glycoprotein contains glycans with negatively charged groups, such as sialic acid and/or sulfate groups; thus, negative ionization mode was used to detect these glycans with higher sensitivity.[36] A representative mass spectrum is in Figure 4.

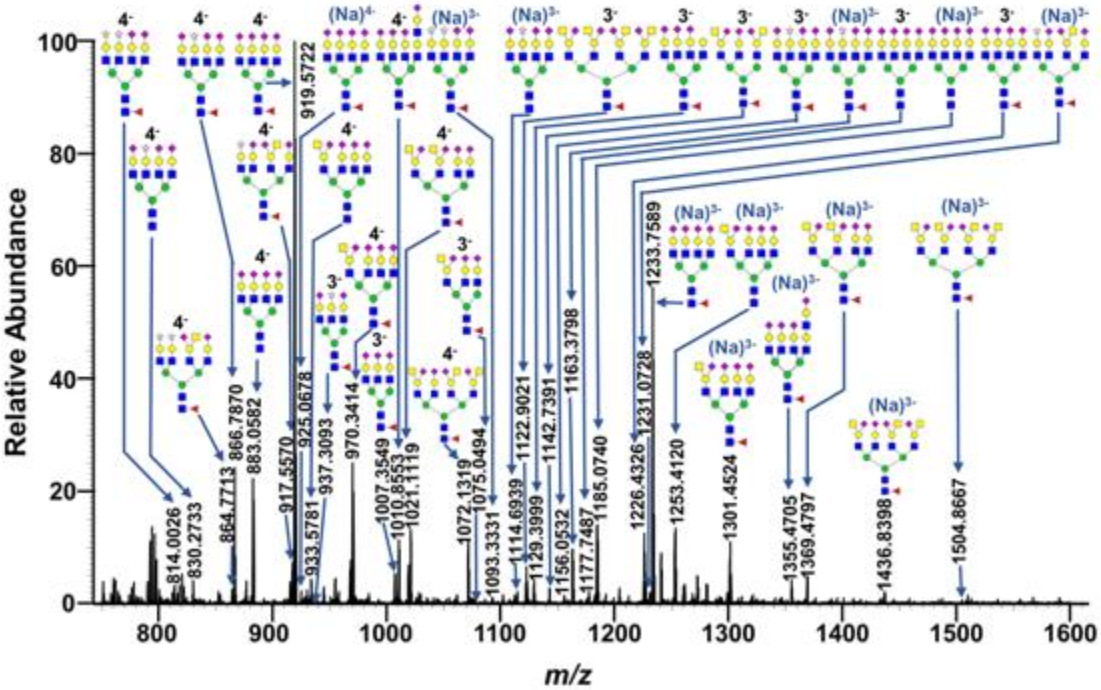
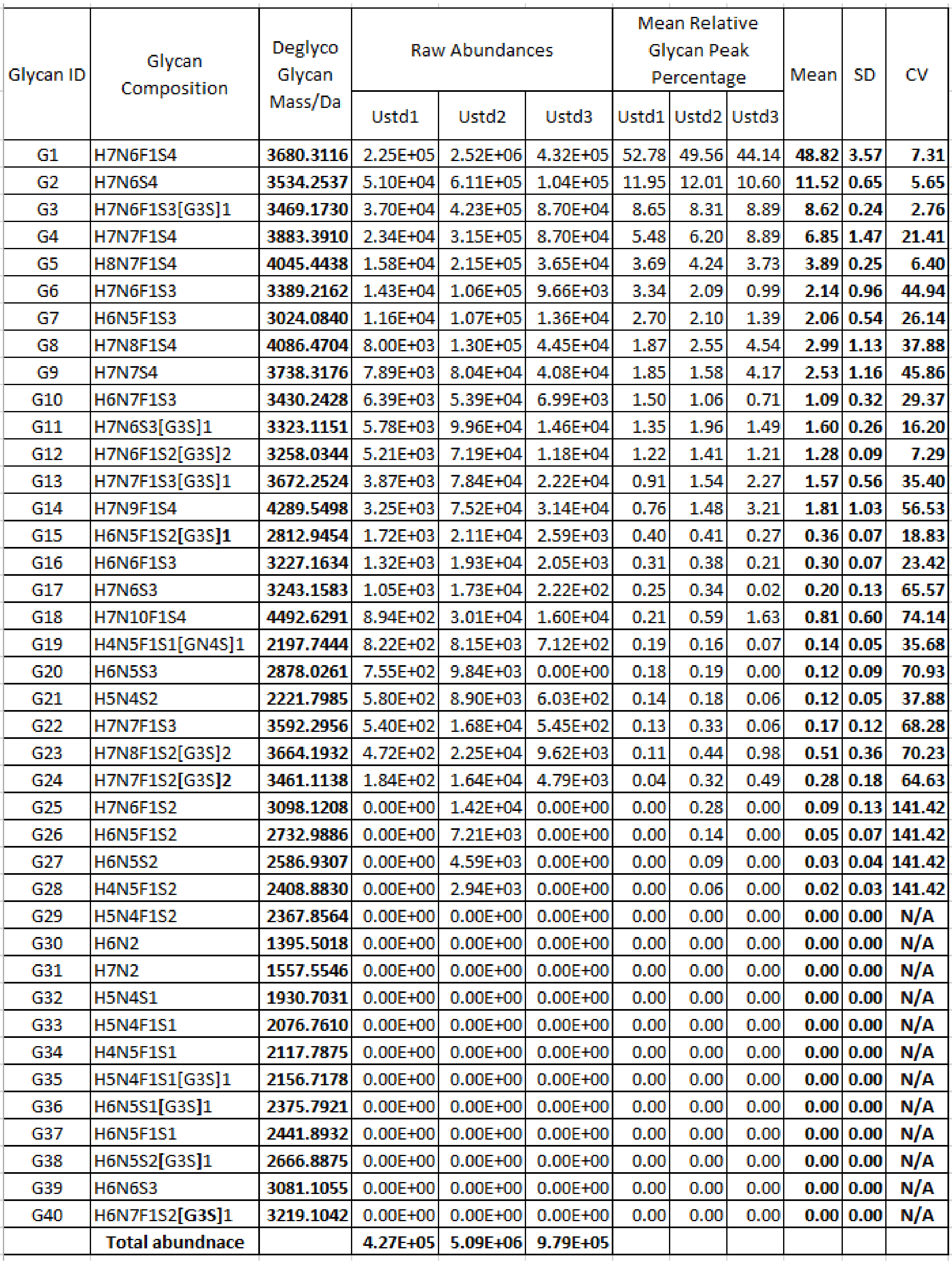


Figure 4. A representative uromodulin *N*-linked glycome profile recorded with direct ESI-MS in negative ion mode. Relative abundance of the eighteen most abundant glycan compositions of the 28 identified glycans are represented. Glycan signals' deprotonated and sodiated adducts ([deglycosylated glycan mass+Na$^+$-4H$^+$]$^{3-}$) are assigned. Monosaccharide units: blue square (*N*-acetylglucosamine), green circle (mannose), yellow circle (galactose), purple diamond (*N*-acetyl neuraminic acid, red triangle (fucose), yellow square (*N*-acetylgalactosamine), and white star (sulfate groups).

The glycans were assigned by comparing the high-resolution MS data to the masses of glycans that had been assigned from uromodulin previously, and assignments were confirmed with MS/MS data as described in the experimental section. This procedure resulted in the assignment of twenty-eight uromodulin glycans of a possible forty that had been reported previously.[5, 9, 12-14, 25]

The developed, direct ESI-MS quantitative approach, accompanied by detection in the negative ion mode, was useful for detecting many of the uromodulin *N*-linked glycans, while omitting any labeling steps. Table 2 includes information of raw abundances and relative glycan peak percentages recorded across individual glycan compositions of uromodulin standards.

Table 2. Represents the raw abundances, mean relative glycan peak percentages and coefficients of variation values calculated for *N*-linked glycans; identified for three uromodulin standards (Ustd1, Ustd2, and Ustd3), derived from three different lots.



Glycan ID	Glycan Composition	Deglyco Glycan Mass/Da	Raw Abundances			Mean Relative Glycan Peak Percentage			Mean	SD	CV
			Ustd1	Ustd2	Ustd3	Ustd1	Ustd2	Ustd3			
G1	H7N6F1S4	**3680.3116**	2.25E+05	2.52E+06	4.32E+05	52.78	49.56	44.14	**48.82**	**3.57**	**7.31**
G2	H7N6S4	**3534.2537**	5.10E+04	6.11E+05	1.04E+05	11.95	12.01	10.60	**11.52**	**0.65**	**5.65**
G3	H7N6F1S3[G3S]1	**3469.1730**	3.70E+04	4.23E+05	8.70E+04	8.65	8.31	8.89	**8.62**	**0.24**	**2.76**
G4	H7N7F1S4	**3883.3910**	2.34E+04	3.15E+05	8.70E+04	5.48	6.20	8.89	**6.85**	**1.47**	**21.41**
G5	H8N7F1S4	**4045.4438**	1.58E+04	2.15E+05	3.65E+04	3.69	4.24	3.73	**3.89**	**0.25**	**6.40**
G6	H7N6F1S3	**3389.2162**	1.43E+04	1.06E+05	9.66E+03	3.34	2.09	0.99	**2.14**	**0.96**	**44.94**
G7	H6N5F1S3	**3024.0840**	1.16E+04	1.07E+05	1.36E+04	2.70	2.10	1.39	**2.06**	**0.54**	**26.14**
G8	H7N8F1S4	**4086.4704**	8.00E+03	1.30E+05	4.45E+04	1.87	2.55	4.54	**2.99**	**1.13**	**37.88**
G9	H7N7S4	**3738.3176**	7.89E+03	8.04E+04	4.08E+04	1.85	1.58	4.17	**2.53**	**1.16**	**45.86**
G10	H6N7F1S3	**3430.2428**	6.39E+03	5.39E+04	6.99E+03	1.50	1.06	0.71	**1.09**	**0.32**	**29.37**
G11	H7N6S3[G3S]1	**3323.1151**	5.78E+03	9.96E+04	1.46E+04	1.35	1.96	1.49	**1.60**	**0.26**	**16.20**
G12	H7N6F1S2[G3S]2	**3258.0344**	5.21E+03	7.19E+04	1.18E+04	1.22	1.41	1.21	**1.28**	**0.09**	**7.29**
G13	H7N7F1S3[G3S]1	**3672.2524**	3.87E+03	7.84E+04	2.22E+04	0.91	1.54	2.27	**1.57**	**0.56**	**35.40**
G14	H7N9F1S4	**4289.5498**	3.25E+03	7.52E+04	3.14E+04	0.76	1.48	3.21	**1.81**	**1.03**	**56.53**
G15	H6N5F1S2[G3S]**1**	**2812.9454**	1.72E+03	2.11E+04	2.59E+03	0.40	0.41	0.27	**0.36**	**0.07**	**18.83**
G16	H6N6F1S3	**3227.1634**	1.32E+03	1.93E+04	2.05E+03	0.31	0.38	0.21	**0.30**	**0.07**	**23.42**
G17	H7N6S3	**3243.1583**	1.05E+03	1.73E+04	2.22E+02	0.25	0.34	0.02	**0.20**	**0.13**	**65.57**
G18	H7N10F1S4	**4492.6291**	8.94E+02	3.01E+04	1.60E+04	0.21	0.59	1.63	**0.81**	**0.60**	**74.14**
G19	H4N5F1S1[GN4S]1	**2197.7444**	8.22E+02	8.15E+03	7.12E+02	0.19	0.16	0.07	**0.14**	**0.05**	**35.68**
G20	H6N5S3	**2878.0261**	7.55E+02	9.84E+03	0.00E+00	0.18	0.19	0.00	**0.12**	**0.09**	**70.93**
G21	H5N4S2	**2221.7985**	5.80E+02	8.90E+03	6.03E+02	0.14	0.18	0.06	**0.12**	**0.05**	**37.88**
G22	H7N7F1S3	**3592.2956**	5.40E+02	1.68E+04	5.45E+02	0.13	0.33	0.06	**0.17**	**0.12**	**68.28**
G23	H7N8F1S2[G3S]2	**3664.1932**	4.72E+02	2.25E+04	9.62E+03	0.11	0.44	0.98	**0.51**	**0.36**	**70.23**
G24	H7N7F1S2[G3S]**2**	**3461.1138**	1.84E+02	1.64E+04	4.79E+03	0.04	0.32	0.49	**0.28**	**0.18**	**64.63**
G25	H7N6F1S2	**3098.1208**	0.00E+00	1.42E+04	0.00E+00	0.00	0.28	0.00	**0.09**	**0.13**	**141.42**
G26	H6N5F1S2	**2732.9886**	0.00E+00	7.21E+03	0.00E+00	0.00	0.14	0.00	**0.05**	**0.07**	**141.42**
G27	H6N5S2	**2586.9307**	0.00E+00	4.59E+03	0.00E+00	0.00	0.09	0.00	**0.03**	**0.04**	**141.42**
G28	H4N5F1S2	**2408.8830**	0.00E+00	2.94E+03	0.00E+00	0.00	0.06	0.00	**0.02**	**0.03**	**141.42**
G29	H5N4F1S2	**2367.8564**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G30	H6N2	**1395.5018**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G31	H7N2	**1557.5546**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G32	H5N4S1	**1930.7031**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G33	H5N4F1S1	**2076.7610**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G34	H4N5F1S1	**2117.7875**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G35	H5N4F1S1[G3S]1	**2156.7178**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G36	H6N5S1[G3S]1	**2375.7921**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G37	H6N5F1S1	**2441.8932**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G38	H6N5S2[G3S]1	**2666.8875**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G39	H6N6S3	**3081.1055**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
G40	H6N7F1S2**[G3S]**1	**3219.1042**	0.00E+00	0.00E+00	0.00E+00	0.00	0.00	0.00	**0.00**	**0.00**	**N/A**
	Total abundnace		**4.27E+05**	**5.09E+06**	**9.79E+05**						

From 40 possible *N*-linked glycans that had been reported for uromodulin previously,[5, 9, 12-14, 25] which are listed in the Table 1 of the Appendix B, a total of 28 glycan compositions were identified for Ustd2 sample (G1-G28), while 24 glycan compositions were identified in both Ustd1 and Ustd3 samples (G1-G24), respectively. Therefore, all the glycans that were identified in at least one of the standard samples were used for the quantitation (G1-G28). Figure 5 illustrates the relative distribution of 28 *N*-linked glycan compositions quantified for the three uromodulin standards.

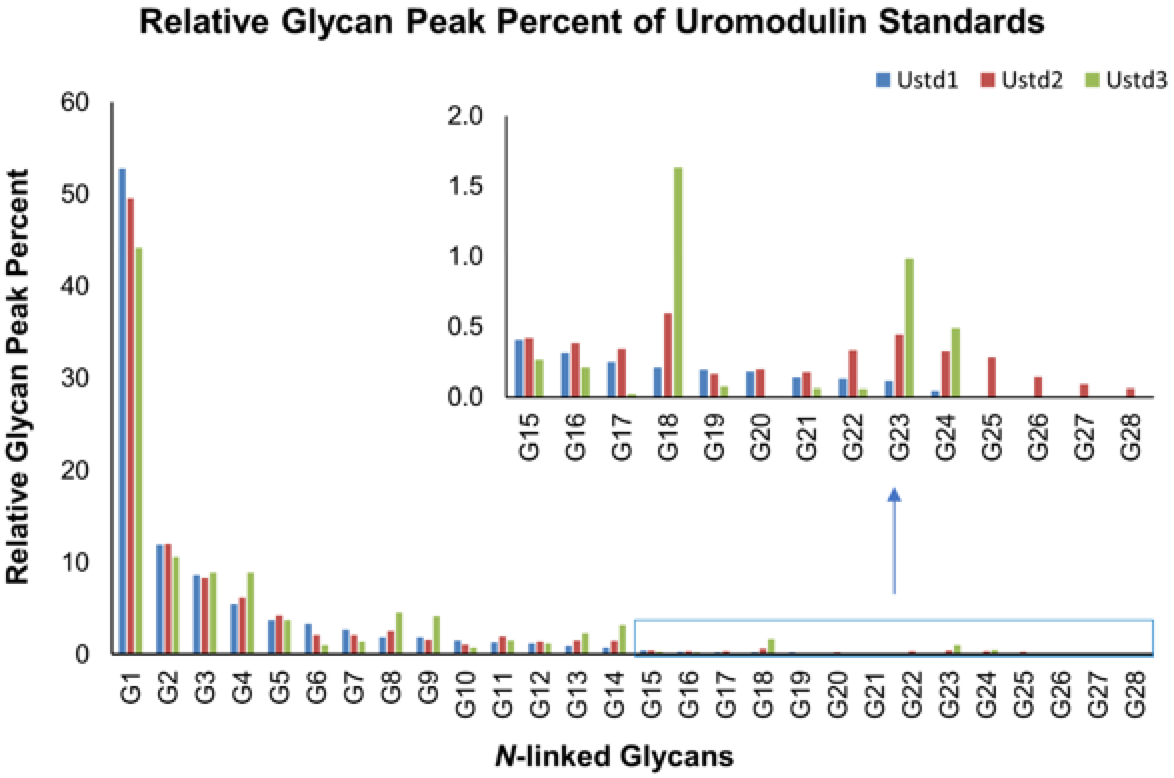


Figure 5. Uromodulin standards' *N*-linked glycans' relative peak percent recorded across 28 glycan compositions. G1 to G28 glycan labeling is in consistent with the *N*-linked glycan list provided in the Table 1 of the Appendix B. Ustd1, Ustd2, and Ustd3 are three different uromodulin standards generated from three different stock solutions; prepared in three different days; analyzed under identical negative ESI-MS conditions. For the clarity of the figure, G15-G28 *N*-linked glycans are zoomed in.

Among the quantified *N*-linked glycans, G1 glycan with H7N6F1S4 composition, which is reported to be a complex-type, tetra-antennary glycan, was the most abundant in all the analyzed uromodulin standard samples, while contributing about 49% to the total glycan pool. Successively, G2 glycan with H7N6S4 composition followed the G1 glycan, while contributing about 12% to the total glycan pool. Among the other glycan compositions quantified for uromodulin standards, 12 glycan compositions (G3 – G14) showed relative glycan peak percentages lower than 10%, but higher than 1%, while the rest of the 14 glycan compositions (G15 – G28) contributed less than 1% to the total glycan pool.

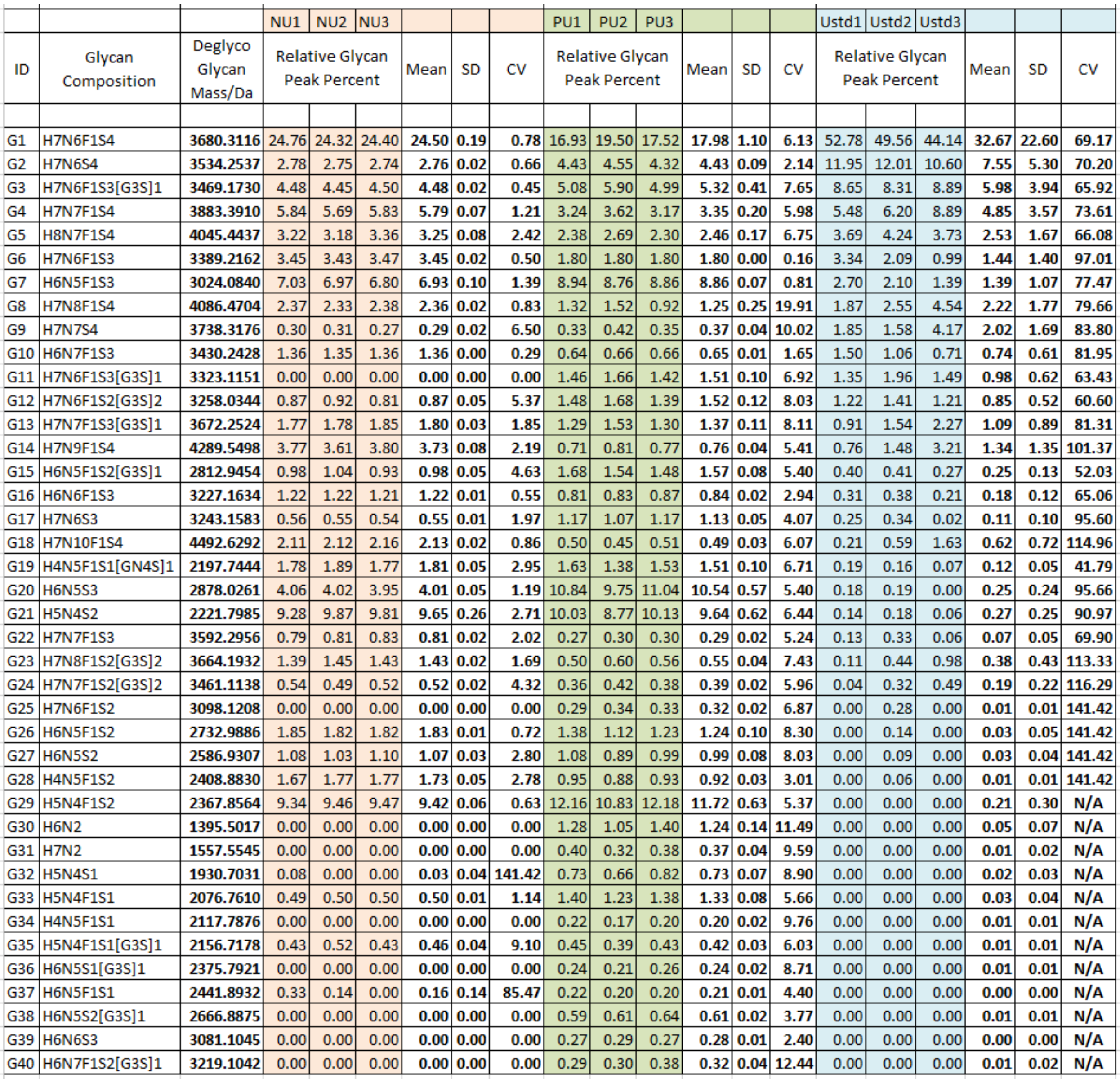
In this study, three different uromodulin standard samples (Ustd1, Ustd2, and Ustd3), derived from three different lots, were analyzed separately to assess lot-to-lot reproducibility of the glycosylation profile. As shown in Table 2, the method yielded less than 8% of coefficients of variation for the relative glycan peak percentages calculated for the three most abundant glycan compositions; these three glycans: G1, G2, and G3 contributed about 49%, 12%, and 8.6% to the total glycan pool. However, the coefficients of variation for all the other glycan compositions, except for G5 and G12, showed relatively higher CV values; this might be a result of lot-to-lot variation of the uromodulin standards used for this study.

3.3.5 Quantification of Uromodulin *N*-linked Glycans Extracted from Human Urine

We next extended the developed approach to analyze *N*-linked glycans of uromodulin; enriched from human urine derived from two different biological states. As uromodulin is the most abundant protein excreted in urine, we performed direct filtration to enrich uromodulin from 10 mL aliquots of urine samples of a pregnant and a normal women. Briefly, urine samples were passed through 50 kD MWCO filters, then, the resulting uromodulin-enriched urinary proteins were desalted by performing multiple washing steps, all prior to the *N*-linked glycan release and quantitation (see Figure 1 for uromodulin isolation and *N*-linked glycan quantitation workflow).

Table 3 includes all the *N*-linked glycans quantified for uromodulin extracted from triplicate samples of a normal urine (NU1, NU2, and NU3) and a pregnant urine sample (PU1, PU2, and PU3), along with three uromodulin standard samples. The method allowed quantitation of a total of 31 and 40 *N*-linked glycan compositions for NU and PU samples, respectively. Similar to the uromodulin standards' quantitation data, in both normal urine and pregnant urine samples; G1 glycan with H7N6F1S4 composition showed to be the highest intense glycan peak; the mean relative peak percentage calculated for G1 glycan was about 49%, 24% and 18% for uromodulin standard, uromodulin, extracted from normal urine, and from pregnant urine samples, respectively. Apart from the most abundant glycan composition, the rest of the glycans quantified for both NU and PU samples contributed less than 12% to the total glycan pool, while about a half of these glycans of each group showed less than 1% of mean relative glycan peak percentages.

Table 3. Mean relative glycan peak percentages and coefficients of variation values calculated for *N*-linked glycans; derived from uromodulin; extracted from triplicate samples of a normal urine (NU1, NU2, and NU3), pregnant urine (PU1, PU2, and PU3), and uromodulin standards (Ustd1, Ustd2, and Ustd3).



			NU1	NU2	NU3				PU1	PU2	PU3				Ustd1	Ustd2	Ustd3			
ID	Glycan Composition	Deglyco Glycan Mass/Da	Relative Glycan Peak Percent			Mean	SD	CV	Relative Glycan Peak Percent			Mean	SD	CV	Relative Glycan Peak Percent			Mean	SD	CV
G1	H7N6F1S4	3680.3116	24.76	24.32	24.40	24.50	0.19	0.78	16.93	19.50	17.52	17.98	1.10	6.13	52.78	49.56	44.14	32.67	22.60	69.17
G2	H7N6S4	3534.2537	2.78	2.75	2.74	2.76	0.02	0.66	4.43	4.55	4.32	4.43	0.09	2.14	11.95	12.01	10.60	7.55	5.30	70.20
G3	H7N6F1S3[G3S]1	3469.1730	4.48	4.45	4.50	4.48	0.02	0.45	5.08	5.90	4.99	5.32	0.41	7.65	8.65	8.31	8.89	5.98	3.94	65.92
G4	H7N7F1S4	3883.3910	5.84	5.69	5.83	5.79	0.07	1.21	3.24	3.62	3.17	3.35	0.20	5.98	5.48	6.20	8.89	4.85	3.57	73.61
G5	H8N7F1S4	4045.4437	3.22	3.18	3.36	3.25	0.08	2.42	2.38	2.69	2.30	2.46	0.17	6.75	3.69	4.24	3.73	2.53	1.67	66.08
G6	H7N6F1S3	3389.2162	3.45	3.43	3.47	3.45	0.02	0.50	1.80	1.80	1.80	1.80	0.00	0.16	3.34	2.09	0.99	1.44	1.40	97.01
G7	H6N5F1S3	3024.0840	7.03	6.97	6.80	6.93	0.10	1.39	8.94	8.76	8.86	8.86	0.07	0.81	2.70	2.10	1.39	1.39	1.07	77.47
G8	H7N8F1S4	4086.4704	2.37	2.33	2.38	2.36	0.02	0.83	1.32	1.52	0.92	1.25	0.25	19.91	1.87	2.55	4.54	2.22	1.77	79.66
G9	H7N7S4	3738.3176	0.30	0.31	0.27	0.29	0.02	6.50	0.33	0.42	0.35	0.37	0.04	10.02	1.85	1.58	4.17	2.02	1.69	83.80
G10	H6N7F1S3	3430.2428	1.36	1.35	1.36	1.36	0.00	0.29	0.64	0.66	0.66	0.65	0.01	1.65	1.50	1.06	0.71	0.74	0.61	81.95
G11	H7N6F1S3[G3S]1	3323.1151	0.00	0.00	0.00	0.00	0.00	0.00	1.46	1.66	1.42	1.51	0.10	6.92	1.35	1.96	1.49	0.98	0.62	63.43
G12	H7N6F1S2[G3S]2	3258.0344	0.87	0.92	0.81	0.87	0.05	5.37	1.48	1.68	1.39	1.52	0.12	8.03	1.22	1.41	1.21	0.85	0.52	60.60
G13	H7N7F1S3[G3S]1	3672.2524	1.77	1.78	1.85	1.80	0.03	1.85	1.29	1.53	1.30	1.37	0.11	8.11	0.91	1.54	2.27	1.09	0.89	81.31
G14	H7N9F1S4	4289.5498	3.77	3.61	3.80	3.73	0.08	2.19	0.71	0.81	0.77	0.76	0.04	5.41	0.76	1.48	3.21	1.34	1.35	101.37
G15	H6N5F1S2[G3S]1	2812.9454	0.98	1.04	0.93	0.98	0.05	4.63	1.68	1.54	1.48	1.57	0.08	5.40	0.40	0.41	0.27	0.25	0.13	52.03
G16	H6N6F1S3	3227.1634	1.22	1.22	1.21	1.22	0.01	0.55	0.81	0.83	0.87	0.84	0.02	2.94	0.31	0.38	0.21	0.18	0.12	65.06
G17	H7N6S3	3243.1583	0.56	0.55	0.54	0.55	0.01	1.97	1.17	1.07	1.17	1.13	0.05	4.07	0.25	0.34	0.02	0.11	0.10	95.60
G18	H7N10F1S4	4492.6292	2.11	2.12	2.16	2.13	0.02	0.86	0.50	0.45	0.51	0.49	0.03	6.07	0.21	0.59	1.63	0.62	0.72	114.96
G19	H4N5F1S1[GN4S]1	2197.7444	1.78	1.89	1.77	1.81	0.05	2.95	1.63	1.38	1.53	1.51	0.10	6.71	0.19	0.16	0.07	0.12	0.05	41.79
G20	H6N5S3	2878.0261	4.06	4.02	3.95	4.01	0.05	1.19	10.84	9.75	11.04	10.54	0.57	5.40	0.18	0.19	0.00	0.25	0.24	95.66
G21	H5N4S2	2221.7985	9.28	9.87	9.81	9.65	0.26	2.71	10.03	8.77	10.13	9.64	0.62	6.44	0.14	0.18	0.06	0.27	0.25	90.97
G22	H7N7F1S3	3592.2956	0.79	0.81	0.83	0.81	0.02	2.02	0.27	0.30	0.30	0.29	0.02	5.24	0.13	0.33	0.06	0.07	0.05	69.90
G23	H7N8F1S2[G3S]2	3664.1932	1.39	1.45	1.43	1.43	0.02	1.69	0.50	0.60	0.56	0.55	0.04	7.43	0.11	0.44	0.98	0.38	0.43	113.33
G24	H7N7F1S2[G3S]2	3461.1138	0.54	0.49	0.52	0.52	0.02	4.32	0.36	0.42	0.38	0.39	0.02	5.96	0.04	0.32	0.49	0.19	0.22	116.29
G25	H7N6F1S2	3098.1208	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.34	0.33	0.32	0.02	6.87	0.00	0.28	0.00	0.01	0.01	141.42
G26	H6N5F1S2	2732.9886	1.85	1.82	1.82	1.83	0.01	0.72	1.38	1.12	1.23	1.24	0.10	8.30	0.00	0.14	0.00	0.03	0.05	141.42
G27	H6N5S2	2586.9307	1.08	1.03	1.10	1.07	0.03	2.80	1.08	0.89	0.99	0.99	0.08	8.03	0.00	0.09	0.00	0.03	0.04	141.42
G28	H4N5F1S2	2408.8830	1.67	1.77	1.77	1.73	0.05	2.78	0.95	0.88	0.93	0.92	0.03	3.01	0.00	0.06	0.00	0.01	0.01	141.42
G29	H5N4F1S2	2367.8564	9.34	9.46	9.47	9.42	0.06	0.63	12.16	10.83	12.18	11.72	0.63	5.37	0.00	0.00	0.00	0.21	0.30	N/A
G30	H6N2	1395.5017	0.00	0.00	0.00	0.00	0.00	0.00	1.28	1.05	1.40	1.24	0.14	11.49	0.00	0.00	0.00	0.05	0.07	N/A
G31	H7N2	1557.5545	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.32	0.38	0.37	0.04	9.59	0.00	0.00	0.00	0.01	0.02	N/A
G32	H5N4S1	1930.7031	0.08	0.00	0.00	0.03	0.04	141.42	0.73	0.66	0.82	0.73	0.07	8.90	0.00	0.00	0.00	0.02	0.03	N/A
G33	H5N4F1S1	2076.7610	0.49	0.50	0.50	0.50	0.01	1.14	1.40	1.23	1.38	1.33	0.08	5.66	0.00	0.00	0.00	0.03	0.04	N/A
G34	H4N5F1S1	2117.7876	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.17	0.20	0.20	0.02	9.76	0.00	0.00	0.00	0.01	0.01	N/A
G35	H5N4F1S1[G3S]1	2156.7178	0.43	0.52	0.43	0.46	0.04	9.10	0.45	0.39	0.43	0.42	0.03	6.03	0.00	0.00	0.00	0.01	0.01	N/A
G36	H6N5S1[G3S]1	2375.7921	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.21	0.26	0.24	0.02	8.71	0.00	0.00	0.00	0.01	0.01	N/A
G37	H6N5F1S1	2441.8932	0.33	0.14	0.00	0.16	0.14	85.47	0.22	0.20	0.20	0.21	0.01	4.40	0.00	0.00	0.00	0.00	0.00	N/A
G38	H6N5S2[G3S]1	2666.8875	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.61	0.64	0.61	0.02	3.77	0.00	0.00	0.00	0.01	0.01	N/A
G39	H6N6S3	3081.1045	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.29	0.27	0.28	0.01	2.40	0.00	0.00	0.00	0.00	0.00	N/A
G40	H6N7F1S2[G3S]1	3219.1042	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.30	0.38	0.32	0.04	12.44	0.00	0.00	0.00	0.01	0.02	N/A

When the quantitation reproducibility of the developed label-free (-)ESI-MS method is considered; the method proved to be highly reproducible for quantifying *N*-linked glycans of uromodulin, extracted from human urine. For both NU and PU sample groups, a very high within-group reproducibility was observed; the CV values obtained for all the *N*-linked glycans of NU sample group were below 9%, except for very low abundant G32 and G37 glycans; and even a lower (<3%) CV's were reported for glycan compositions that contributed more than 1% to the total glycan pool. Similarly, for pregnant urine triplicates, less than 13% of CV values were reported for all of 40 *N*-linked glycans, except for one glycan composition. Overall, these results clearly showed that the presented approach is highly reproducible while showing its applicability towards targeted quantification of *N*-linked glycans derived from complex biological matrices.

3.3.6 Quantification of Glycosylation Differences of Three Sample Groups

During this study, *N*-linked glycosylation profiles of three uromodulin standards, uromodulin, extracted from triplicate samples of a normal urine and a pregnant urine sample were quantified; Figure 6 represents the relative distribution of uromodulin *N*-linked glycans derived from triplicates of these three sample groups.

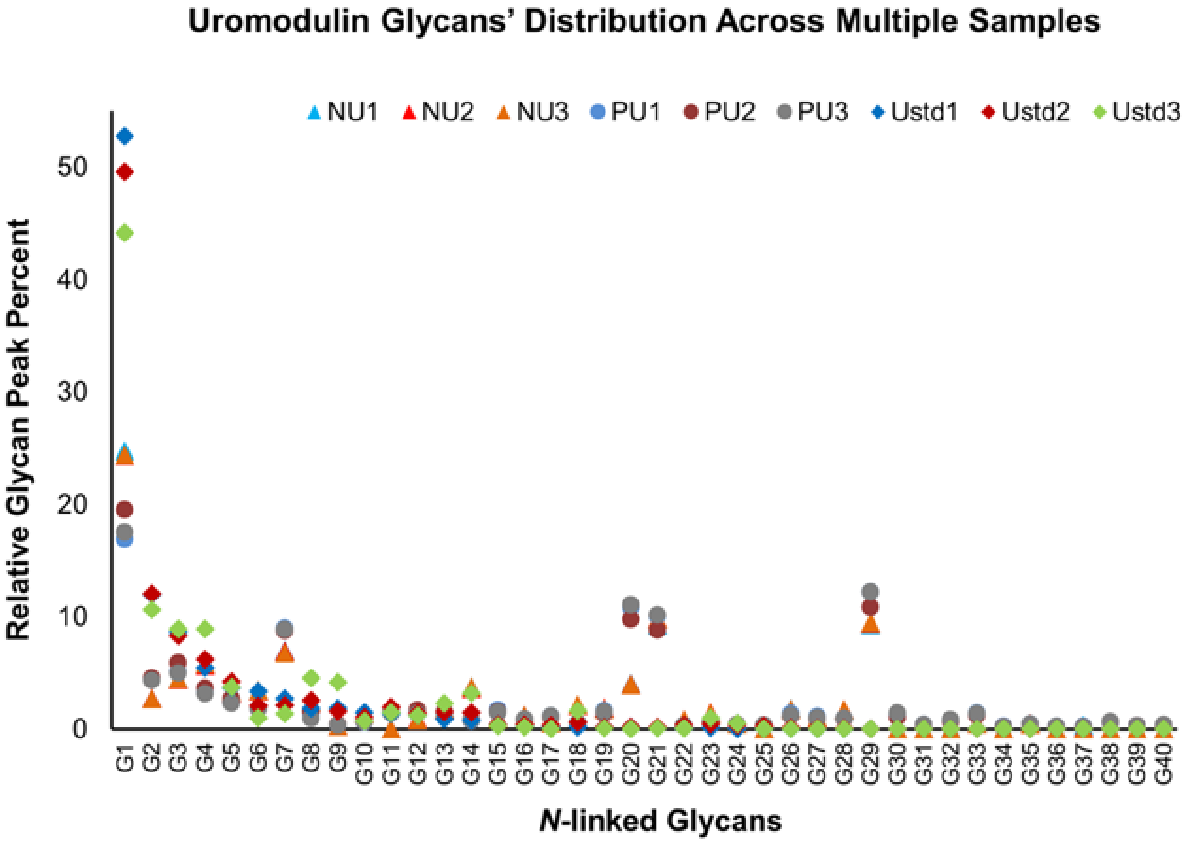


Figure 6. *N*-linked glycans' Relative peak percentages reported for three separate uromodulin standards (Ustd1, Ustd2, Ustd3), uromodulin, extracted from triplicate samples of a normal urine (NU1, NU2, NU3) and a pregnant urine sample (PU1, PU2, PU3). Glycosylation differences observed in each group were very subtle; however, the data showed high within-group reproducibility.

Among the quantified *N*-linked glycans, many of the glycans showed subtle glycosylation differences in each group; however, the method allowed quantitation of these subtle changes while showing a very high within-group reproducibility. For an example, G22-G24 glycans in NU samples contributed less than 1.5% to the total glycan pool, but the within group reproducibility obtained for these three glycans were very high, while showing CV values below 5%. Similarly, these three glycans contributed less than 0.6% to the total glycan pool of the PU samples; but, still the within-group reproducibility was high and reported CV values less than

8%. This example, along with the data presented in the Table 3, clearly show that the developed method provides highly reproducible data over multiple sample preparations and over multiple analyses.

Figure 7 represents PCA data generated for three different sample groups; normal urine (group 1), pregnant urine (group 2), and uromodulin standards (group 3).

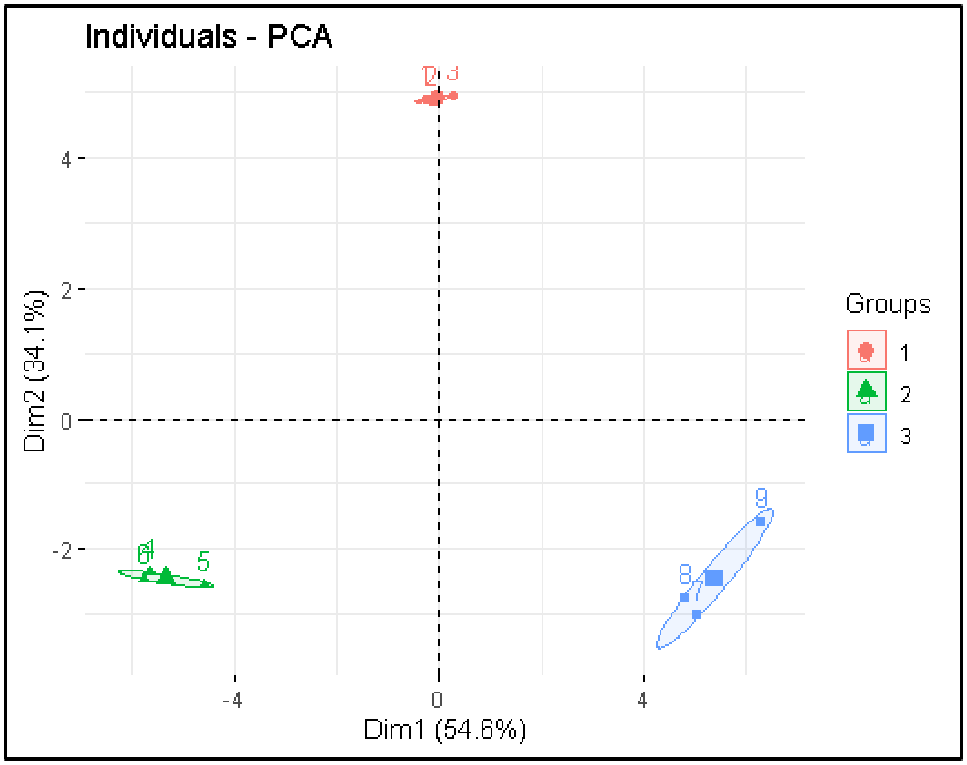


Figure 7. Classification of three sample groups based on their glycosylation data by using Principal component analysis (PCA). Group 1, group 2, and group 3 represent samples of normal urine (NU), pregnant urine (PU), and uromodulin standards (Ustd), respectively. The data clearly showed that all three sample groups are unique, even though the glycan differences within them are subtle.

These data clearly show that these three groups are unique and clearly separable, even though the glycosylation differences observed among the groups are subtle. For both group 1 and group 2 samples, within-group clustering was very tight; this is because of the high reproducibility provided by the developed approach. Even for the group 3, variability within the group was not very broad. The greater spread in this group was very likely to be a result of lot-to-lot variability of the uromodulin standards, yet, this variability is very small compared to the biological variability among the three groups.

3.4 Conclusion

We developed a rapid, direct ESI-MS approach to quantify *N*-linked glycans that ionize well in the negative ion mode. The method is straightforward, omits any glycan labeling steps, which typically require additional post-sample clean up steps prior to the analysis. The developed (-)ESI-MS method was applied to quantify *N*-linked glycans of standard glycoproteins; it proved to be highly reproducible across multiple sample preparations and multiple analyses. Then, we further extended this method to quantify *N*-linked glycans of uromodulin, directly extracted from human urine samples of two different biological states; the observed glycosylation differences were subtle in each group; however, within-group reproducibility provided by the method was very high. Moreover, all of the analyzed samples were clearly separable into distinct, sample-related groups, even though the glycosylation differences among groups were subtle. Therefore, this method can be applied in quantitative glycomics studies, as it is a simple, straightforward one, which effectively permits highly reproducible quantitation data even though the glycosylation differences of the glycomics samples are subtle.

3.5 Acknowledgements

This work was funded by NIH grant R35GM130354 to Heather Desaire. We are truly thankful for their generous support.

3.6 References

1. Liu, S.; Cheng, L.; Fu, Y.; Liu, B. F.; Liu, X., *J Proteomics* **2018,** *181*, 225-237.

2. Jia, X.; Chen, J.; Sun, S.; Yang, W.; Yang, S.; Shah, P.; Hoti, N.; Veltri, B.; Zhang, H., *Proteomics* **2016,** *16* (23), 2989-2996.

3. Kamiyama, T.; Yokoo, H.; Furukawa, J.; Kurogochi, M.; Togashi, T.; Miura, N.; Nakanishi, K.; Kamachi, H.; Kakisaka, T.; Tsuruga, Y.; Fujiyoshi, M.; Taketomi, A.; Nishimura, S.; Todo, S., *Hepatology (Baltimore, Md.)* **2013,** *57* (6), 2314-25.

4. Tan, Z.; Yin, H.; Nie, S.; Lin, Z.; Zhu, J.; Ruffin, M. T.; Anderson, M. A.; Simeone, D. M.; Lubman, D. M., *Journal of proteome research* **2015,** *14* (4), 1968-78.

5. Argade, S.; Chen, T.; Shaw, T.; Berecz, Z.; Shi, W.; Choudhury, B.; Parsons, C. L.; Sur, R. L., *Urolithiasis* **2015,** *43* (4), 303-12.

6. Vivekanandan-Giri, A.; Slocum, J. L.; Buller, C. L.; Basrur, V.; Ju, W.; Pop-Busui, R.; Lubman, D. M.; Kretzler, M.; Pennathur, S., *International journal of proteomics* **2011,** *2011,* 214715.

7. Reiding, K. R.; Bondt, A.; Hennig, R.; Gardner, R. A.; O'Flaherty, R.; Trbojevic-Akmacic, I.; Shubhakar, A.; Hazes, J. M. W.; Reichl, U.; Fernandes, D. L.; Pucic-Bakovic, M.; Rapp, E.; Spencer, D. I. R.; Dolhain, R.; Rudd, P. M.; Lauc, G.; Wuhrer, M., *Molecular & cellular proteomics : MCP* **2019,** *18* (1), 3-15.

8. Russell, A. C.; Simurina, M.; Garcia, M. T.; Novokmet, M.; Wang, Y.; Rudan, I.; Campbell, H.; Lauc, G.; Thomas, M. G.; Wang, W., *Glycobiology* **2017,** *27* (5), 501-510.

9. van Rooijen, J. J.; Voskamp, A. F.; Kamerling, J. P.; Vliegenthart, J. F., *Glycobiology* **1999,** *9* (1), 21-30.

10. Argade, S. P.; Vanichsarn, C.; Chenoweth, M.; Parsons, C. L., *BJU international* **2009,** *103* (8), 1085-9.

11. Bates, J. M.; Raffi, H. M.; Prasadan, K.; Mascarenhas, R.; Laszik, Z.; Maeda, N.; Hultgren, S. J.; Kumar, S., *Kidney international* **2004,** *65* (3), 791-7.

12. Parsons, C. L.; Stein, P.; Zupkas, P.; Chenoweth, M.; Argade, S. P.; Proctor, J. G.; Datta, A.; Trotter, R. N., *The Journal of urology* **2007,** *178* (6), 2665-70.

13. Hard, K.; Van Zadelhoff, G.; Moonen, P.; Kamerling, J. P.; Vliegenthart, F. G., *European journal of biochemistry* **1992,** *209* (3), 895-915.

14. van Rooijen, J. J.; Kamerling, J. P.; Vliegenthart, J. F., *European journal of biochemistry* **1998,** *256* (2), 471-87.

15. Serafini-Cessi, F.; Bellabarba, G.; Malagolini, N.; Dall'Olio, F., *Journal of immunological methods* **1989,** *120* (2), 185-9.

16. Deckert, T.; Kofoed-Enevoldsen, A.; Vidal, P.; Nørgaard, K.; Andreasen, H. B.; Feldt-Rasmussen, B., *Diabetologia* **1993,** *36* (3), 244-51.

17. Kanauchi, M.; Nishioka, H.; Hashimoto, T.; Dohi, K., *Nihon Jinzo Gakkai shi* **1995,** *37* (11), 649-54.

18. Talks, B. J.; Bradwell, S. B.; Delamere, J.; Rayner, W.; Clarke, A.; Lewis, C. T.; Thomas, O. D.; Bradwell, A. R., *High altitude medicine & biology* **2018,** *19* (3), 295-298.

19. Song, T.; Aldredge, D.; Lebrilla, C. B., *Anal Chem* **2015,** *87* (15), 7754-62.

20. Sun, S.; Hu, Y.; Ao, M.; Shah, P.; Chen, J.; Yang, W.; Jia, X.; Tian, Y.; Thomas, S.; Zhang, H., *Clinical proteomics* **2019,** *16*, 35.

21. Shah, B.; Jiang, X. G.; Chen, L.; Zhang, Z., *Journal of the American Society for Mass Spectrometry* **2014,** *25* (6), 999-1011.

22. Stadlmann, J.; Pabst, M.; Kolarich, D.; Kunert, R.; Altmann, F., *Proteomics* **2008,** *8* (14), 2858-71.

23. Zhang, Y.; Go, E. P.; Desaire, H., *Anal Chem* **2008,** *80* (9), 3144-58.

24. Hong, C.; Abdullah, M.; Wong, N., *Int J Pharm Pharm Sci* **2013,** *5* (3), 385-89.

25. Smagula, R. M.; Van Halbeek, H.; Decker, J. M.; Muchmore, A. V.; Moody, C. E.; Sherblom, A. P., *Glycoconj J* **1990,** *7* (6), 609-24.

26. Rebecchi, K. R.; Wenke, J. L.; Go, E. P.; Desaire, H., *Journal of the American Society for Mass Spectrometry* **2009,** *20* (6), 1048-59.

27. Jackson, A. U.; Talaty, N.; Cooks, R. G.; Van Berkel, G. J., *Journal of the American Society for Mass Spectrometry* **2007,** *18* (12), 2218-25.

28. Karki, S.; Shi, F.; Archer, J. J.; Sistani, H.; Levis, R. J., *Journal of the American Society for Mass Spectrometry* **2018,** *29* (5), 1002-1011.

29. Gonzalez-Dominguez, R.; Castilla-Quintero, R.; Garcia-Barrera, T.; Gomez-Ariza, J. L., *Analytical biochemistry* **2014,** *465*, 20-7.

30. Johannesson, N.; Pearce, E.; Dulay, M.; Zare, R. N.; Bergquist, J.; Markides, K. E., *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **2006,** *842* (1), 70-4.

31. Wei, L.; Cai, Y.; Yang, L.; Zhang, Y.; Lu, H., *Anal Chem* **2018,** *90* (17), 10442-10449.

32. Tsai, T. H.; Wang, M.; Di Poto, C.; Hu, Y.; Zhou, S.; Zhao, Y.; Varghese, R. S.; Luo, Y.; Tadesse, M. G.; Ziada, D. H.; Desai, C. S.; Shetty, K.; Mechref, Y.; Ressom, H. W., *Journal of proteome research* **2014,** *13* (11), 4859-68.

33. Yang, S.; Wang, M.; Chen, L.; Yin, B.; Song, G.; Turko, I. V.; Phinney, K. W.; Betenbaugh, M. J.; Zhang, H.; Li, S., *Scientific reports* **2015,** *5*, 17585.

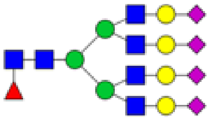
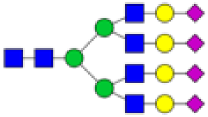
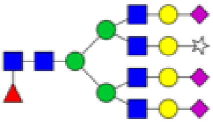
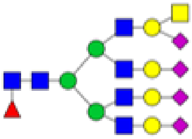
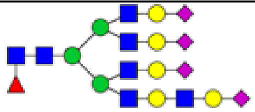
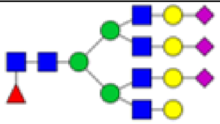
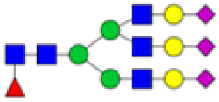
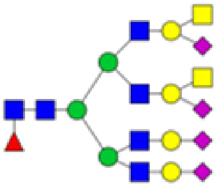
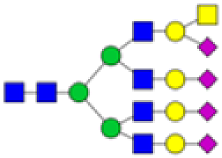
34. Zhou, S.; Hu, Y.; Veillon, L.; Snovida, S. I.; Rogers, J. C.; Saba, J.; Mechref, Y., *Anal Chem* **2016,** *88* (15), 7515-22.

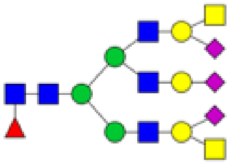
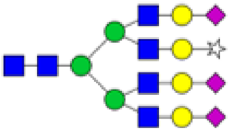
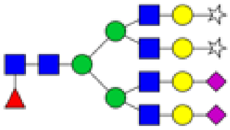
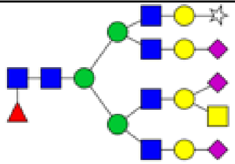
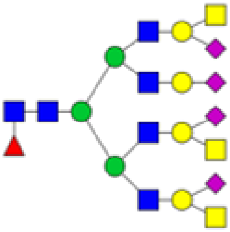
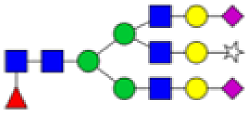
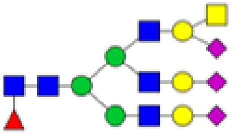
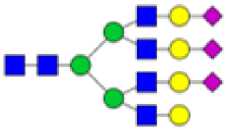
35. Hong, Q.; Ruhaak, L. R.; Stroble, C.; Parker, E.; Huang, J.; Maverakis, E.; Lebrilla, C. B., *Journal of proteome research* **2015,** *14* (12), 5179-92.

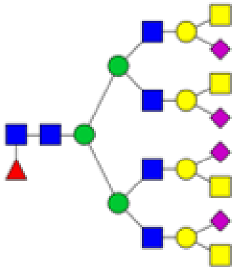
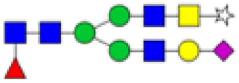
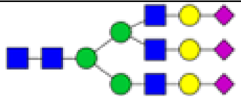
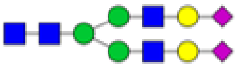
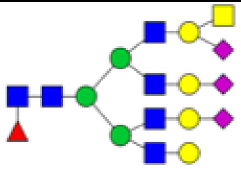
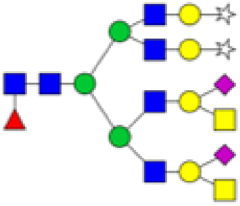
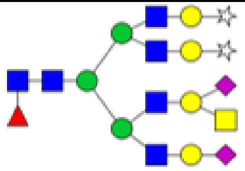
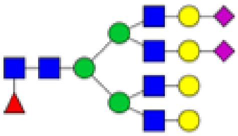
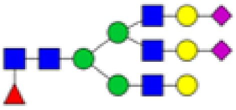
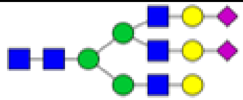
36. Ni, W.; Bones, J.; Karger, B. L., *Anal Chem* **2013,** *85* (6), 3127-35.

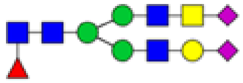
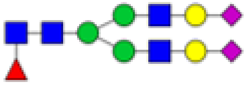
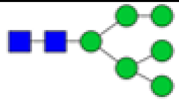
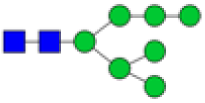
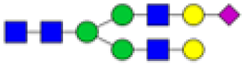
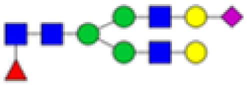
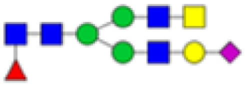
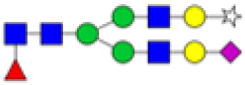
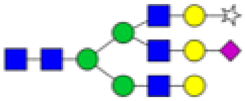
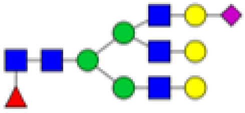
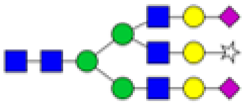
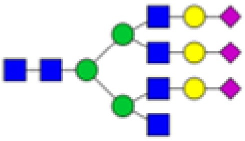
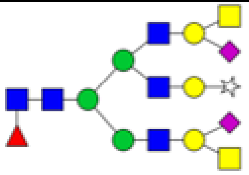
3.7 Appendix B

Table 1. The list of 40 potential *N*-linked glycans prepared for uromodulin glycoprotein in consistent with the literature.[5, 9, 12-14, 25] These compositions were confirmed by using MS/MS data as described in the data analysis section.

Glycan ID	Glycan composition	Glycan structure	Deglycosylated mass/Da
G1	H7N6F1S4		3680.3116
G2	H7N6S4		3534.2537
G3	H7N6F1S3[G3S]1		3469.1730
G4	H7N7F1S4		3883.3910
G5	H8N7F1S4		4045.4438
G6	H7N6F1S3		3389.2162
G7	H6N5F1S3		3024.0840
G8	H7N8F1S4		4086.4704
G9	H7N7S4		3738.3176

Glycan ID	Glycan composition	Glycan structure	Deglycosylated mass/Da
G10	H6N7F1S3		3430.2428
G11	H7N6S3[G3S]1		3323.1151
G12	H7N6F1S2[G3S]2		3258.0344
G13	H7N7F1S3[G3S]1		3672.2524
G14	H7N9F1S4		4289.5498
G15	H6N5F1S2[G3S]1		2812.9454
G16	H6N6F1S3		3227.1634
G17	H7N6S3		3243.1583

Glycan ID	Glycan composition	Glycan structure	Deglycosylated mass/Da
G18	H7N10F1S4		4492.6291
G19	H4N5F1S1[GN4S]1		2197.7444
G20	H6N5S3		2878.0261
G21	H5N4S2		2221.7985
G22	H7N7F1S3		3592.2956
G23	H7N8F1S2[G3S]2		3664.1932
G24	H7N7F1S2[G3S]2		3461.1138
G25	H7N6F1S2		3098.1208
G26	H6N5F1S2		2732.9886
G27	H6N5S2		2586.9307

Glycan ID	Glycan composition	Glycan structure	Deglycosylated mass/Da
G28	H4N5F1S2		2408.8830
G29	H5N4F1S2		2367.8564
G30	H6N2		1395.5018
G31	H7N2		1557.5546
G32	H5N4S1		1930.7031
G33	H5N4F1S1		2076.7610
G34	H4N5F1S1		2117.7875
G35	H5N4F1S1[G3S]1		2156.7178
G36	H6N5S1[G3S]1		2375.7921
G37	H6N5F1S1		2441.8932
G38	H6N5S2[G3S]1		2666.8875
G39	H6N6S3		3081.1055
G40	H6N7F1S2[G3S]1		3219.1042

Chapter 4. Chemically Generated Large Sets of IgG Glycopeptides' Liquid Chromatography-Mass Spectrometry (LC-MS) Data for the Optimization of a Classifier that Uses Whole Glycomic Profile to Classify Samples into Disease versus Healthy

A part of the data described in this chapter has been published in Analytical Chemistry: David Hua, Milani Wijeweera Patabandige, Eden P. Go and Heather Desaire. Anal. Chem. ***2019****, 91, 17, 11070-11077. It is reproduced here with the permission from The American Chemical Society.*

Abstract

Classification of glycomics samples by considering the whole glycomic profile, instead of selecting a single or a few glycan features, may be more useful in tracking the underlying trends of the glycomics data to effectively classify samples into their accurate groups. To test this hypothesis, the Aristotle Classifier; a newly developed classification algorithm, which uses all the individual glycan abundances and their relative proportions to each other to classify samples, was recently developed. Once the classifier was built, it needed to be optimized with challenging glycomics data; however, acquiring clinical glycomics samples from diseased patients and healthy controls, where known glycosylation differences differentiated the sample sets, was a challenge. Therefore, we generated large sets of glycomics data in-house, to represent samples of two biologically different states as healthy and disease; these samples were prepared by slightly altering the glycosylation profile of human Immunoglobulin G (IgG) glycoprotein. Sample preparations were optimized to generate two groups of IgG glycopeptides samples in which

healthy state represented native IgG glycosylation profile while the disease state represented slightly altered IgG glycosylation (sialylation or fucosylation) profile. Once the LC-MS glycomics data were generated; they were analyzed with the Aristotle Classifier, in addition to a standard classification system known as PCA (Principal Components Analysis), to compare the classification capabilities of each method. The Aristotle Classifier outperformed the sample classification of the standard approach while showing its capability of accurately classifying glycomics samples; therefore, this new classification algorithm is indeed useful for the clinical biomarker field.

4.1 Introduction

Protein glycosylation analyses provide unique opportunities in the biomarker discovery field because glycosylated proteins are subjected to change during the progression of many diseases including various cancers,[1-8] Parkinson's disease,[9] heart disease,[10] kidney diseases,[11-12] and Alzheimer's disease.[13] The analyses of protein glycosylation is important especially to diagnose and monitor diseases; thus, specific glycans[1, 5-11, 13] or glycopeptides,[2-3] which are over/under expressed and/or altered, could be used to distinguish patients with various diseases from that of healthy individuals.

Glycomics data differ from proteomics or genomics data because of the very heterogeneous nature of glycans; this heterogeneity is a result of non-template driven and enzymatically-controlled, glycans' biosynthesis process.[14-15] Therefore, glycans' heterogeneity typically complicates the glycomics analyses because of the splitting of the glycosylation signal of any given protein into many different protein glycoforms, which are usually presented in low abundance.[16] However, these diverse glycans/glycopeptides are useful in biomarker discovery field, because they provide multiple features that can be used to distinguish samples of a healthy state from that of a disease state. When these different glycans are used to classify samples into disease versus healthy; the standard approaches typically use one glycan signature[17-18] or a few glycan signatures[5, 8] that best discriminate samples into their accurate groups; but, these methods omit considering all the glycans or glycopeptides in a sample, which can be useful in providing important information that might improve the sample classification.

The concept of using whole glycomic profile of a sample, instead of using one or a few glycan features to indicate a disease state is well explained in a recent report [19] by using an artistic analogy. Briefly, in a fragmented image, as shown in Figure 1A, all the individual pieces, when viewed one at a time, do a sub-optimal job in showing the underlying object that they represent; by analogy, this is similar considering isolated glycan features that could be potential biomarkers, to classify a disease state sample. In contrast, if we bring some of the fragmented pieces of the image together while viewing them in context to each other; part of the image can be built, as shown in Figure 1B; but still, it is not sufficient to provide complete information of the representative image. Likewise, glycans also can be viewed in context to each other by comparing glycan peak ratios of a sample instead of their absolute abundances. Throughout the literature, the use of a single or a few glycan peak ratios as potential biomarkers, to classify samples into disease versus healthy, is reported; however, considering of "thousands" of possible glycan ratios, as a whole signal or a single scorable unit, is certainly new in the biomarker discovery field. This is analogous to bringing all the fragmented pieces of the image together while viewing them in context to each other to build the entire image, as shown in Figure 1C, which is the best in classifying the underlying information of the image of the interest; which indeed shows that the image is about a bird sitting on a branch.

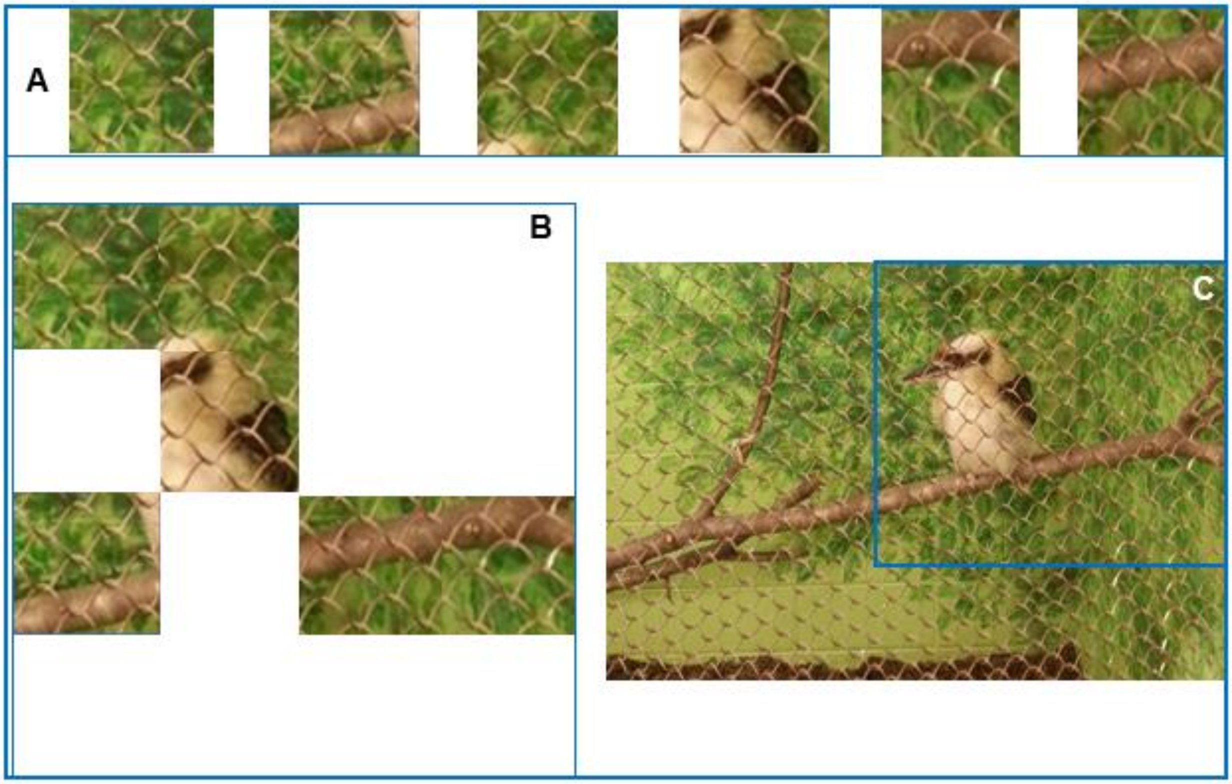


Figure 1. Panel **(A)** represents six separate image fragments, which, individually, are not useful at defining the image of the bird. **(B)** The same six separate image fragments arranged by viewing them in context to each other, which in turn, more useful in classifying the image of the bird. **(C)** The entire image of the bird sitting on a branch, which provides the most useful information for identification of the object.

Based on the concept described above, our group has developed a new, supervised machine learning algorithm, known as the Aristotle Classifier, which uses not only all the glycans, but also their ratios with each other for classification of glycomics samples. The methods section of the reference 19 contains detailed information about the Aristotle Classifier, including feature building, discriminating feature identification, and scoring for sample classification. Once the classifier was built, a set of data was required to optimize it; however, identifying two sets of clinical glycomics data with known differences in their glycosylation profiles was a challenge. Therefore, the intellectual contribution to the Aristotle classifier project undertaken as part of this dissertation, was to chemically generate two groups of glycomics data to mimic two different biological states, healthy versus disease, by slightly altering the glycosylation profiles of a model glycoprotein, human immunoglobulin G or IgG.

IgG is by far the most abundant glycoprotein in human plasma and serum;[1] it has four subclasses as IgG1, IgG2, IgG3, and IgG4, each represents approximately about 60%, 32%, 4%, and 4% relative abundances, respectively.[20] This glycoprotein has a single *N*-linked glycosylation site, located at the N297 position of the CH2 domain of the fragment crystalizable (Fc) region[9, 21] and it is occupied by many different *N*-linked glycans, majorly complex-type bi-antennary glycans, which can be fucosylated (major type), sialylated, bi-sected or not.[9, 22] These different glycans present on IgG introduce huge diversity to its glycosylation profile and they can affect the solubility, stability and therapeutic activity of recombinantly expressed biotherapeutics.[23-24] Not only that, alteration to these glycans, especially, the degree of sialylation, fucosylation, or galactosylation,[1-2, 9, 25] have been reported during the progression of many diseases; including various type of cancers,[1-2, 25] autoimmune diseases,[2] and Parkinson's disease;[9] thus, they serve as potential biomarkers.

Herein we chemically generated large sets of LC-MS data for IgG glycopeptides; these glycopeptides were generated to mimic two different biological states as healthy versus disease. The IgG glycopeptide data of healthy group represented a native IgG glycosylation profile while the disease group represented IgG glycopeptides those were purposely altered either by slightly changing their degree of sialylation or the fucosylation. Multiple samples belonging to healthy versus disease states were chemically generated; they were analyzed under identical LC-MS conditions, followed by quantitation of different glycoforms of IgG1 and IgG2 based on their high-resolution MS signal. Finally, the generated quantitation data were submitted to the Aristotle Classifier to test its ability to classify these chemically generated large sets of IgG glycopeptides data based on their whole glycomic profile. The generated data were useful; the data system successfully classified many of the IgG glycopeptide samples into their accurate groups, far better than the standard classification system in the current field, the Principal Components Analysis (PCA) and showed that the developed classifier can be successfully applied to classify glycomics samples of different biological states.

4.2 Experimental

4.2.1 Materials and Reagents

Human serum IgG, ammonium bicarbonate, guanidine hydrochloride (GdnHCl), dithiothreitol (DTT), iodoacetamide (IAM), formic acid and HPLC grade acetonitrile and methanol were purchased from Sigma Aldrich (St. Louis, MO). Sequencing grade trypsin was from Promega (Madison, WI), and α2-3,6,8,9 Neuraminidase A (Sialidase A), α1-2,3,4,6 fucosidase, 10X glycobuffer (pH 5.5), 100X BSA, was from New England BioLabs (Ipswich, MA). Ultrapure water was obtained from a Direct-Q water purification system (MilliporeSigma, Darmstadt, Germany).

4.2.2 Preparation of Native and Partially Desialylated IgG Tryptic Digests

Two aliquots of Human IgG samples; each containing 100 µg of glycoprotein dissolved in 10X glycobuffer at pH 5.5, were treated separately with either the α2-3,6,8,9 Neuraminidase A enzyme (2.0 µL) or with an identical volume of deionized water. Both samples were incubated for 1 week at 37 °C, the samples' pH was adjusted to pH 8.0 with 300 mM NH_4OH followed by diluting the samples with 50 mM ammonium bicarbonate buffer (pH 8.0) to obtain IgG glycoprotein solutions with a 4 mg/mL final concentration. To denture each glycoprotein sample, GdnHCl was added separately to give 6 M final concentration. Then, to reduce the disulfide bonds, DTT was added to the glycoprotein solutions to a 10 mM final concentration, followed by sample incubation at room temperature for 1 h. Thereafter, disulfide bonds were alkylated by adding IAM to a final concentration of 25 mM, and this reaction was carried out in the dark, at room temperature for 1 h. After the alkylation step, the excess IAM was neutralized by adding DTT to the reaction mixture (at a 30 mM final concentration), and the reaction was continued for 30 mins at room temperature. Then, the resultant glycoprotein solutions were filtered separately through 10 kD MWCO filters; subjected to buffer exchange twice, followed by diluting the resultant glycoprotein concentrate with NH_4HCO_3 buffer (50 mM, pH 8.0) to yield 1 µg/µL concentrated glycoprotein solutions. Thereafter, the trypsin digestion was performed by adding trypsin to each glycoprotein solution at a protein-to-enzyme ratio of 30:1, followed by incubating at 37 °C for 20 hours. Finally, the trypsin digestion was stopped by adding 1 µL of formic acid to each 100 µL of glycoprotein solution and the resultant IgG tryptic digests were aliquoted and stored at -20 °C until the analysis is performed.

4.2.3 Preparation of Native and Partially Defucosylated IgG Tryptic Digests

IgG glycoprotein (160 µg) was dissolved in 50 mM NH_4HCO_3 buffer at pH 8.0, to give a 4 mg/mL concentrated glycoprotein solution; then, the glycoprotein solution was denatured by adding GdnHCl (at 6 M final concentration). The denatured glycoprotein was then reduced with DTT and alkylated with IAM; excess IAM was neutralized with DTT to yield final solution concentrations as similar to the previous section; the added reagent volumes were adjusted based on the initial glycoprotein amount. After these steps, the resultant glycoprotein solution was filtered through a 10 kD MWCO filter and was buffer exchanged two times with the NH_4HCO_3 buffer at pH 8.0. Subsequently, the glycoprotein concentrate was collected through reverse spin (1000 g × 2 min) and diluted with the buffer to give a 1 µg/µL final concentration prior to the trypsin digestion. Then, trypsin was added to the glycoprotein solution at a protein-to-enzyme ratio of 30:1 and incubated for 20 h at 37 °C. After the trypsin digestion, the pH of the IgG tryptic digest was adjusted to pH 5.5 by using 0.01% formic acid; then, the tryptic digest was filtered through 10 kD MWCO filters to remove trypsin, and the filtrate was collected. The filtrate that contains a mixture of IgG glycopeptides and peptides was aliquoted into two fractions; both aliquots (67 µL each) were treated with equal volumes (7.6 µL of each) of 10X glycobuffer and 10X BSA, which was diluted from 100X BSA stock solution. Then α1-2,3,4,6 fucosidase enzyme (10 µL) was added to one sample aliquot to obtain defucosylated IgG, while the other aliquot was treated with an equal volume of 10X glycobuffer to obtain a native or a control sample. After that, both the aliquots were incubated at 37 °C for 1 week; filtered through 10 kD MWCO filters, separately, to remove BSA and/or fucosidase enzyme, and then, the filtrates were collected. Subsequently, the final volumes of the filtrates were brought up to 80 µL, after acidifying them with the 0.1% FA. Both IgG glycopeptide samples: native and partially defucosylated, were then stored at -20 °C prior to the analysis.

4.2.4 IgG Mixed Sample Preparation with Partially Desialylated or Partially Defucosylated IgG Samples

To prepare various IgG native samples mixed with IgG desialo- or defuco- samples, first, tryptic digests of partially desialylated or partially defucosylated IgG samples, which were at 1.0 µg/µL initial concentration were diluted five times with deionized water to obtain 0.2 µg/µL concentrated IgG tryptic samples, separately. Subsequently, IgG desialo 0%, 5%, and 20% mixed samples were prepared; 0% sample was prepared by directly diluting appropriate volume of IgG native sialo control sample (at 1.0 µg/µL) with deionized water, to yield 0.5 µg/µL final concentration; then, 5% and 20% IgG desialo mixed samples were prepared by mixing appropriate volumes of IgG native sialo control sample (1.0 µg/µL) and partially desialylated

IgG sample (0.2 µg/µL), while maintaining the final sample concentrations of all the samples at 0.5 µg/µL. A similar approach was used to prepare IgG defuco 0%, 5%, and 20% mixed samples by mixing appropriate volumes of IgG native fuco control sample and the partially defucosylated IgG sample, at a fixed final concentration (0.5 µg/µL). All the samples were run in triplicates.

4.2.5 IgG Sialo, IgG Desialo 20% Mixed and IgG Fuco, IgG Defuco 20% Mixed Samples Preparation

IgG sialo native and IgG desialo 20% mixed samples were prepared by using native and partially desialylated IgG tryptic digest samples, which were at 0.9 µg/µL initial concentration. From these samples, native IgG tryptic peptides sample at 0.05 µg/µL was prepared by diluting the original native IgG sialo sample (0.9 µg/µL) to obtain stock solution 1 at 0.45 µg/µL concentration, followed by subsequent dilution of stock 1 with appropriate volume of deionized water. This sample is known as IgG sialo sample or group 1. Prior to the preparation of IgG desialo 20% mixed sample, partially desialylated IgG tryptic peptide sample at 0.9 µg/µL concentration was diluted with deionized water to obtain stock solution 2 at 0.1 µg/µL of concentration. Then, IgG desialo 20% mixed sample was prepared by mixing appropriate volumes of stock 1 and stock 2, while maintaining the final concentration of the sample at 0.05 µg/µL. This sample is referred to as IgG desialo 20% mixed sample or group 2.

In addition to group 1 and group 2 sample preparation, IgG fuco native and IgG defuco 20% mixed samples were also prepared, separately. IgG fuco native sample (group 3) at 0.1 µg/µL final concentration was prepared by diluting 0.9 µg/µL concentrated IgG native tryptic digest with deionized water, directly. Then, IgG partially defucosylated tryptic digest solution at 0.9 µg/µL concentration was diluted three times with deionized water to obtain a stock solution at 0.3 µg/µL. Then, appropriate volumes of this stock solution (0.3 µg/µL) and IgG native tryptic digest (at 0.9 µg/µL) was mixed to generate IgG 20% defuco mixed sample, while keeping its final concentration at 0.1 µg/µL. This sample is henceforth referred to as group 4.

4.2.6 Liquid Chromatography-Mass Spectrometry Analysis of IgG Glycopeptide Samples

IgG glycopeptide samples were separated in a reverse phase C18 capillary column (3.5 µm, 300 µm i.d. ×10 cm, Agilent Technologies, Santa Clara, CA) connected online to a Waters Acquity high performance liquid chromatography system (Milford, MA) followed by mass spectrometric (MS) data acquisition using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose, CA). For each run, 3 µL of sample volume was injected into the C18 column with a mobile phase flow rate of 10 µL/min. A gradient elution was performed to separate IgG glycopeptides with mobile phase A and mobile phase B; mobile phase A consists of 99.9% of water with 0.1% formic acid while the mobile phase B consists of 99.9% acetonitrile with 0.1% of formic acid. The liquid chromatography (LC) gradient used for the study was as follows. The column was equilibrated by running 5% of mobile phase B for 3 mins, followed by linear increase of B from 5% to 20% in 22 min to separate the glycopeptides. Then B was ramped to 90% in 20 min for glycopeptide elution, followed by decrease of B to 5% in 5 min, and re-equilibrating the column at 5% B for another 10 mins. During the IgG sample runs, when IgG tryptic samples prepared in sections 4.2.2, 4.2.3, and 4.2.4 were subjected to LC separation, blank runs were performed in between each sample run to minimize the sample carryover. However, when IgG native and IgG desialo/defuco 20% mixed samples described in the section 4.2.5 were subjected to LC separation, blank runs were performed either in between each sample or after two sample runs, as described next.

For IgG glycopeptides' large data set generation, the samples prepared as described in section 4.2.5 were used. During the sample runs, the samples from either the group 1 and group2 or group 3 and group 4 were alternated during the LC-MS analyzes. For group 1 (IgG sialo native) and group 2 (IgG desialo 20% mixed), a total of 21 samples were acquired for each group in two different days: during the first day, altogether 12 samples were acquired, which included 6 samples from each group; 3 weeks after, the second data set with 30 samples, which included 15 samples per each group, were acquired. When acquiring the small data set, blank runs were included in between each sample run, for large data set, a single blank run was included after each pair of sample runs. For group 3 (IgG fuco native) and group 4 (IgG defuco 20% mixed) samples, small data set with five sample runs for each group were acquired in the first day by including blank runs in between each sample run. The large data set for group 3 and group 4 samples were acquired after 3 weeks, 14 sample runs were included for each group, and blank runs were performed after a pair of sample runs.

4.2.7 Mass Spectrometry (MS) Conditions

Electrospray ionization (ESI)-MS in the positive ion mode with a heated ion source, which was held at 2.3 kV was used. The temperature of the ion transfer tube and the vaporizer was set as 300 ˚C and 20 ˚C, respectively. Full MS scans were acquired with the Orbitrap resolution at 60 k (at *m/z* 200) and the scan range was set at *m/z* range of 400 – 2000. The AGC target and the maximum ion injection time were set at 4×10^5 and 50 ms, respectively. Data dependent MS/MS data were acquired to confirm the glycopeptide compositions; collision-induced dissociation (CID) data were collected by selecting the first five most abundant peaks from the full MS run. CID spectra were collected in the ion trap with a rapid scan rate, exclusion duration was set at 30 s with a repeated count of one. For CID, AGC target of 2×10^3 and maximum injection time of 300 ms was used. Furthermore, during the MS/MS data acquisition, 2 Da isolation width was used for parent ions selection, and the selected precursor ions were fragmented by applying 35% of collision energy for 10 ms. Once the full MS and CID data were analyzed and IgG glycopeptides were confidently identified; fifteen different IgG1glycopeptides and thirteen different IgG2 glycopeptides for group 1 and 2 analysis, thirteen IgG2 glycopeptides for group 3 and 4 analysis, were identified for the quantification. For glycopeptide quantitation, raw abundances of each glycoform recorded at 50% abundance of extracted ion chromatogram (XIC) were recorded; then, the relative glycan peak percentage for each individual glycopeptide was calculated by dividing the individual glycopeptide's raw abundance from the total abundance calculated by summing the raw abundances of all the fifteen IgG1 or thirteen IgG2 glycopeptides (for group 1 and 2) or all thirteen IgG2 glycopeptides (for group 3 and 4) recorded for each sample run. Appendix C, Table 1, Table 2, and Table 3 show the relative glycan peak percentages calculated across IgG1 and IgG2 glycopeptides generated from 42 samples of group 1 and 2 pair and IgG2 glycopeptides generated from 38 samples of group 3 and 4 pair, respectively.

4.2.8 Classification of Sample Groups

The quantitation data generated for IgG glycopeptides identified for group 1 and 2 and group 3 and 4 samples were separately submitted for data classification by Principal Components Analysis (PCA) and the Aristotle Classifier. PCA was conducted in R, version 3.5.2, using the package "factoextra". The data were centered and scaled prior to the PCA transformation. The Aristotle Classifier was performed using the version of the software provided in reference 19; it was run in R, version 3.5.0.

4.3 Results and Discussion

4.3.1 Overview of the Study

Glycomics sample classification based on the whole glycomic profile of a sample, including all the individual abundances of samples' glycans/glycopeptides, along with their relative proportions to each other, is certainly new in the glycomics analysis field, but this approach could be helpful for identifying useful trends in the glycomics data. The Aristotle Classifier is a new data system developed by our group; its classification is based on using whole glycomic profile of a sample as a single scorable unit, instead of using a single feature or a few features to classify the samples. Once the new classifier was built, a challenge data set of mass spectrometry data from glycomics samples, which could enable the optimization of the classifier, was required; thus, large sets of IgG glycopeptides data were generated in-house, to mimic two different biological states as healthy and disease. Among these samples, samples with a native IgG glycopeptide profile were considered as a healthy state, while the IgG samples with slightly altered sialylation or fucosylation were considered as the disease state. However, generating IgG glycopeptide samples to mimic two different biological states was a challenge, and it required careful selection and optimization of both the sample preparation and the instrument conditions. Thus, during this study, first, IgG glycopeptides were generated, LC-MS methods were optimized for effective separation of multiple IgG subclasses and their glycopeptides, followed by identification of IgG glycosylation profile. After that, methods were optimized to generate partially deglycosylated (partially desialylated or defucosylated) IgG tryptic digests; these samples were used to slightly contaminate the glycosylation profile of a native IgG tryptic digest in order to generate samples of a disease state. Finally, large sets of LC-MS IgG glycopeptides data were generated by using IgG tryptic digests with native glycosylation profile (healthy state) and that with a slightly altered glycosylation profile (disease state). The resulted IgG glycopeptides' data were quantified and submitted to the newly developed Aristotle Classifier and also to another standard classifier (PCA), to challenge their classification ability. Most of the generated data proved to be optimum, and these data were classified more accurately by the new classification system compared to the standard approach.

4.3.2 Identification of IgG Glycopeptides

Immunoglobulin G (IgG), the most abundant glycoprotein in human plasma has four sub classes as IgG1, IgG2, IgG3, and IgG4, each represents approximately 60%, 32%, 4%, and 4% relative abundances, respectively.[20] Each sub class of IgG has a single glycosylation site majorly occupied by complex-type bi-antennary glycans that can be fucosylated (major type), sialylated, bisected, or not.[9, 22] Therefore, optimization of LC-MS conditions were necessary not only to effectively separate these different IgG subclasses, but also to obtain better ionization for multiple IgG glycopeptides. Figure 2A and 2B represent two extracted ion chromatograms (XICs) obtained for triply charged IgG1 and IgG2 peptides, each attached to a bi-antennary, core-fucosylated, [Hex]4[HexNAc]4[Fuc]1 glycan composition. As shown in this Figure, we were able to detect and separate the two most abundant subclasses of IgG: IgG1 and IgG2, each bears peptide sequence of EEQYNSTYR (mw =1188.5047) and EEQFNSTFR (mw=1156.5149), respectively. Two well separated chromatographic peaks were obtained for IgG1 and IgG2 glycopeptides, and they eluted approximately around 15.26 min and 21.54 min, respectively.

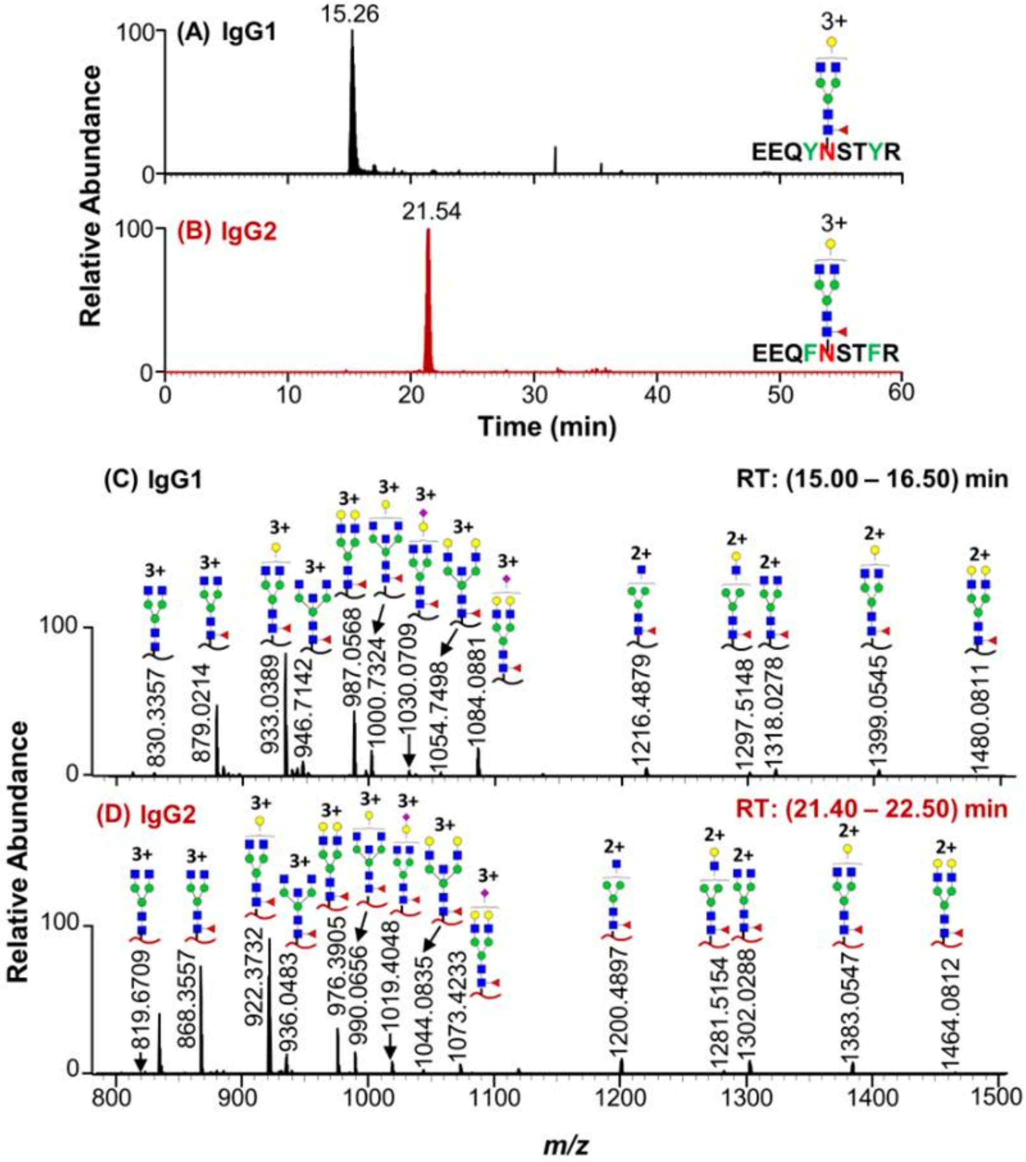
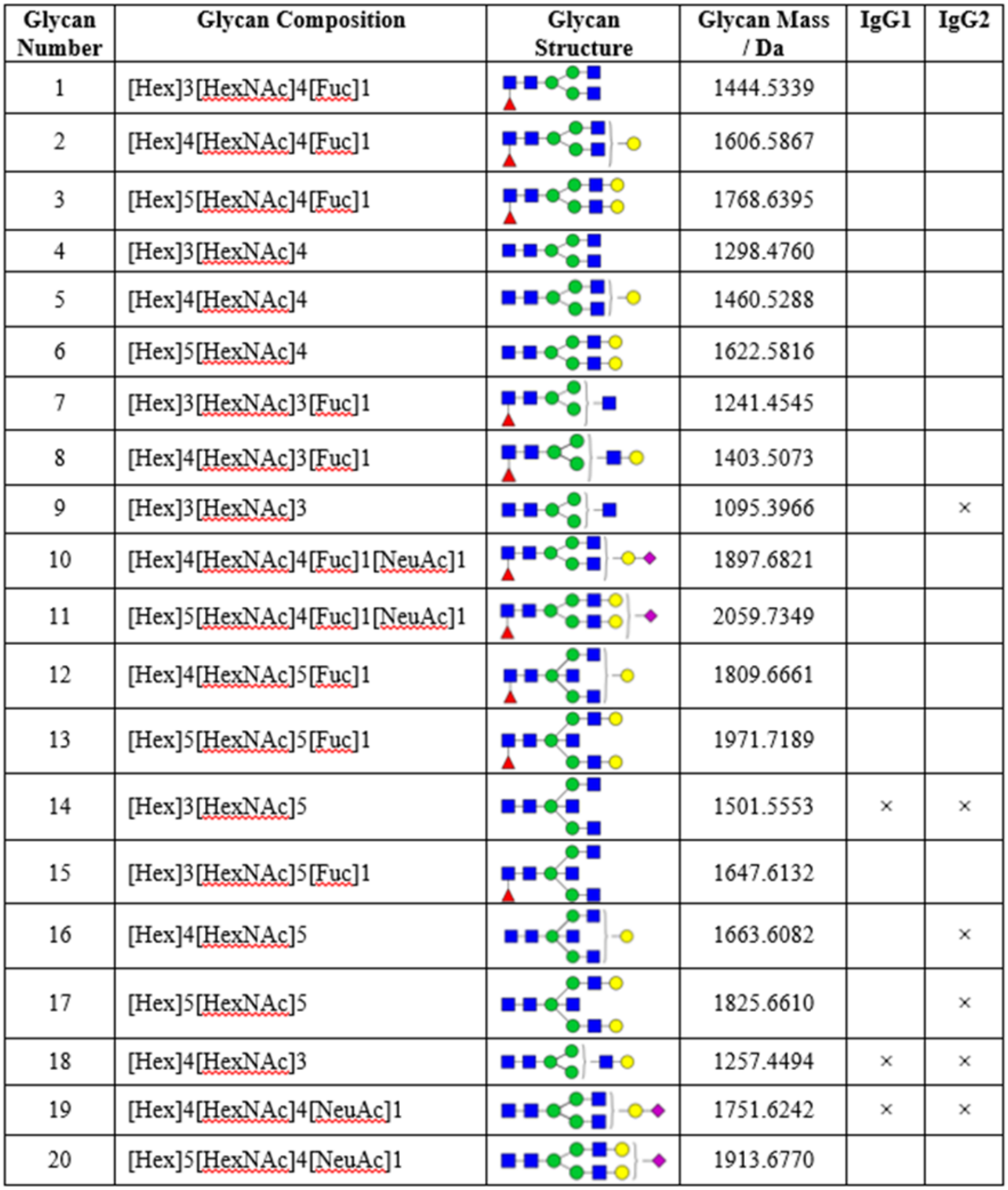


Figure 2. Extracted ion chromatograms (XICs) obtained for triply charged IgG1 (EEQYNSTYR) **(A)** and IgG2 (EEQFNSTFR) **(B)** peptides attached to [Hex]4[HexNAc]4[Fuc]1 glycan composition. IgG1 and IgG2 showed well separated peaks in the same LC-MS run while showing an approximate retention time of 15.26 min (IgG1) and 21.54 min (IgG2), respectively. Figure 2 **(C)** and **(D)** represent *N*-linked glycopeptides assigned for IgG1 (15.00 - 16.50 min) and IgG2 (21.40 – 22.50 min), across their multiple charge states. Monosaccharide units: blue square (*N*-acetylglucosamine), green circle (mannose), yellow circle (galactose), purple diamond (*N*-acetyl neuraminic acid), and red triangle (fucose).

For each identified IgG sub class, possible glycopeptides were searched against a list of potential *N*-linked glycans that had been assigned from IgG in previous reports;[16, 26] glycan compositions were assigned by comparing the high-resolution MS data to the theoretical *m/z*'s of the glycans within 5 ppm mass error, and these assignments were then confirmed by using CID data. Figure 2C and 2D represent *N*-linked glycopeptide profiles obtained for IgG1 and IgG2 glycopeptides for the retention time ranges of (15.00 -16.50) min and (21.40 – 22.50) min, respectively. Table 1 represents the *N*-linked glycans identified for both IgG1 and IgG2 after performing multiple analyses on tryptic digests of native IgG samples. In addition, this list includes some other possible *N*-linked glycans that are expected to be observed when the glycosylation changes are introduced to the IgG glycosylation profile, as described in sections 4.2.4 and 4.2.5.

Table 1. The list of 20 *N*-linked glycans identified for IgG1 and IgG2 glycopeptides. Among these glycans, those labeled with the crossed mark (×) were not observed in native IgG samples being analyzed, but they were expected to be observed when the glycosylation changes are introduced to the IgG glycosylation profile.



Glycan Number	Glycan Composition	Glycan Structure	Glycan Mass / Da	IgG1	IgG2
1	[Hex]3[HexNAc]4[Fuc]1		1444.5339		
2	[Hex]4[HexNAc]4[Fuc]1		1606.5867		
3	[Hex]5[HexNAc]4[Fuc]1		1768.6395		
4	[Hex]3[HexNAc]4		1298.4760		
5	[Hex]4[HexNAc]4		1460.5288		
6	[Hex]5[HexNAc]4		1622.5816		
7	[Hex]3[HexNAc]3[Fuc]1		1241.4545		
8	[Hex]4[HexNAc]3[Fuc]1		1403.5073		
9	[Hex]3[HexNAc]3		1095.3966		×
10	[Hex]4[HexNAc]4[Fuc]1[NeuAc]1		1897.6821		
11	[Hex]5[HexNAc]4[Fuc]1[NeuAc]1		2059.7349		
12	[Hex]4[HexNAc]5[Fuc]1		1809.6661		
13	[Hex]5[HexNAc]5[Fuc]1		1971.7189		
14	[Hex]3[HexNAc]5		1501.5553	×	×
15	[Hex]3[HexNAc]5[Fuc]1		1647.6132		
16	[Hex]4[HexNAc]5		1663.6082		×
17	[Hex]5[HexNAc]5		1825.6610		×
18	[Hex]4[HexNAc]3		1257.4494	×	×
19	[Hex]4[HexNAc]4[NeuAc]1		1751.6242	×	×
20	[Hex]5[HexNAc]4[NeuAc]1		1913.6770		

4.3.3 Quantifying the Degree of IgG Deglycosylation Achieved at Glycopeptide Level

The ultimate goal of this study was to generate two groups of IgG samples to mimic two biologically different states as healthy and disease; these two groups were expected to generate by using a tryptic digest of native IgG as the healthy state, and a tryptic digest of IgG with slightly altered glycosylation as the disease state. Altered glycosylation of IgG is associated with the progression of many diseases, and some of these alterations include changes to the degree of sialylation, fucosylation, and galcatosylation.[1-2, 9, 25] Therefore, a disease state sample was expected to generate by slightly adulterating the native glycosylation profile of a IgG tryptic digest with a partially desialylated or a partially defucosylated IgG tryptic digest. Therefore, as the first step of the sample preparation, partially desialylated or partially defucosylated IgG samples needed to be prepared; the detailed protocols are described in the experimental sections 4.2.2 and 4.2.3.

Figure 3A and 3B show representative workflows of the implemented sample preparation protocol to generate tryptic digests of partially desialylated and partially defucosylated IgG samples along with their control samples, respectively.

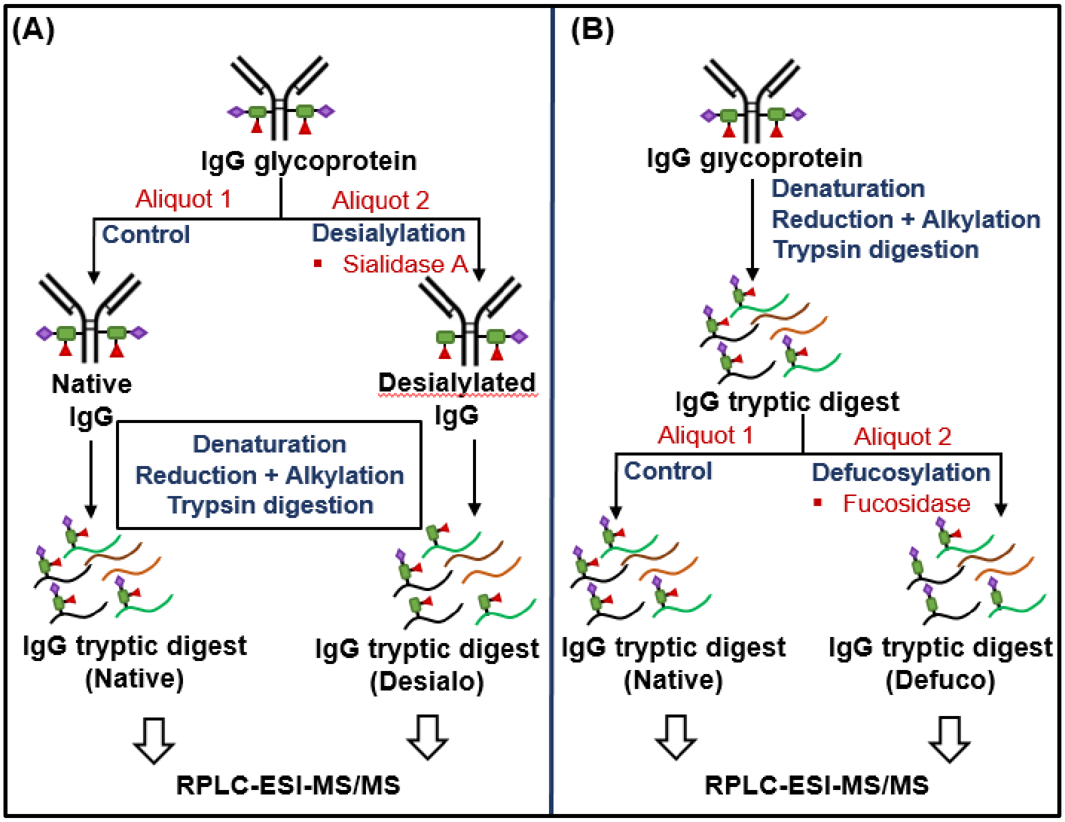


Figure 3. Representative workflows that used to generate tryptic digests of partially desialylated IgG and its control sample (IgG native tryptic digest) **(A)** and tryptic digest of partially defucosylated IgG and its control sample (IgG native tryptic digest) **(B)**. The resulted IgG tryptic digest samples were analyzed under identical LC-ESI-MS/MS conditions and the data were used to calculate the degree of desialylation or defucosylation achieved with the optimized sample preparation protocol.

In these two sample preparation protocols, the sialidase A reaction was performed at the glycoprotein level, while the fucosidase reaction was carried out at the glycopeptide level to improve the digestion efficiency of the fucosidase enzyme. After obtaining LC-MS data for IgG desialylated/defucosylated tryptic digest samples and the control samples (IgG native tryptic digests); the relative glycopeptide peak percentages of each individual glycopeptide, from both sample types, were calculated; then, the degree of desialylation/defucosylation achieved at the individual glycopeptide level was calculated as follows.

$$Deglycosylation\ \% = \frac{Relative\ glycopeptide\ peak\ percent\ (deglycosylated\ sample - native\ sample)}{Relative\ glycopeptide\ peak\ percent\ of\ native\ sample}$$

Figure 4 represents the relative glycopeptide peak percentages calculated for 14 different *N*-linked IgG2 glycopeptides generated from IgG native and IgG desialo tryptic digests.

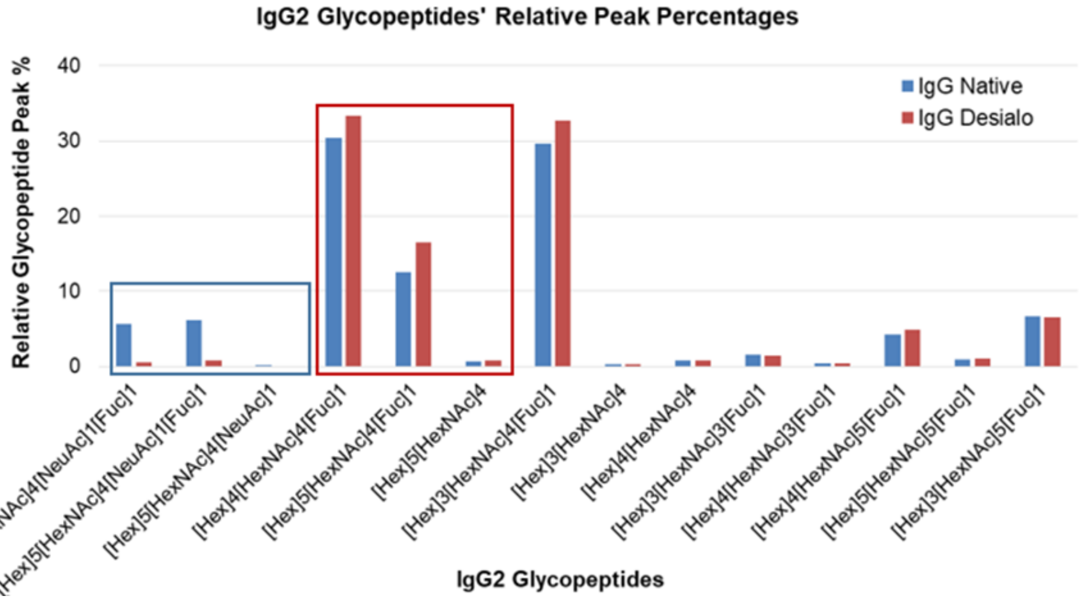


Figure 4. Relative peak percentages calculated for IgG2 glycopeptides generated from IgG native and IgG desialo tryptic digest samples.

In this figure, the glycan compositions in the blue box represent three sialylated glycans, the red box includes three non-sialylated glycans that correspond to the sialylated ones, while the rest of the eight glycan compositions represent other non-sialylated glycans identified in IgG2 glycopeptides. After treating a IgG sample with sialidase A enzyme; upon completion of the reaction, in an ideal situation, three observation were expected in the IgG desialo sample as compared to the IgG native sample; these observations include; (1) a decrease of relative peak percentages of sialylated glycopeptides, (2) an increase of relative peak percentages of corresponding non-sialylated glycopeptides, and (3) unchanged relative peak percentages for the rest of the non-sialylated glycopeptides. As shown in the Figure 4, the relative peak percentages of individual glycopeptides of IgG desialo digest, as compared to the IgG native digest, were changed; these changes included a (88 – 100) % decrease of relative glycopeptide peak percent of sialylated glycopepetides, (10 – 31) % increase of relative glycopeptide peak percentages of corresponding desialylated glycopepetides and less than 18% relative glycopeptide percentage differences for rest of the glycopeptides. This result clearly shows that the optimized desialylation protocol generated partially desialylated IgG2 glycopeptides that showed about 88% to 100% of desialylation at the individual glycopeptide levels.

Similarly, IgG2 glycopeptides' relative peak percentages obtained for the IgG defucosylated sample compared to the IgG native sample were used to calculate the degree of defucosylation achieved by following the defucosylation protocol shown in Figure 3B. The calculated relative peak percentages of thirteen IgG2 glycopeptides identified in IgG native and IgG defuco tryptic digested samples are shown in Figure 5.

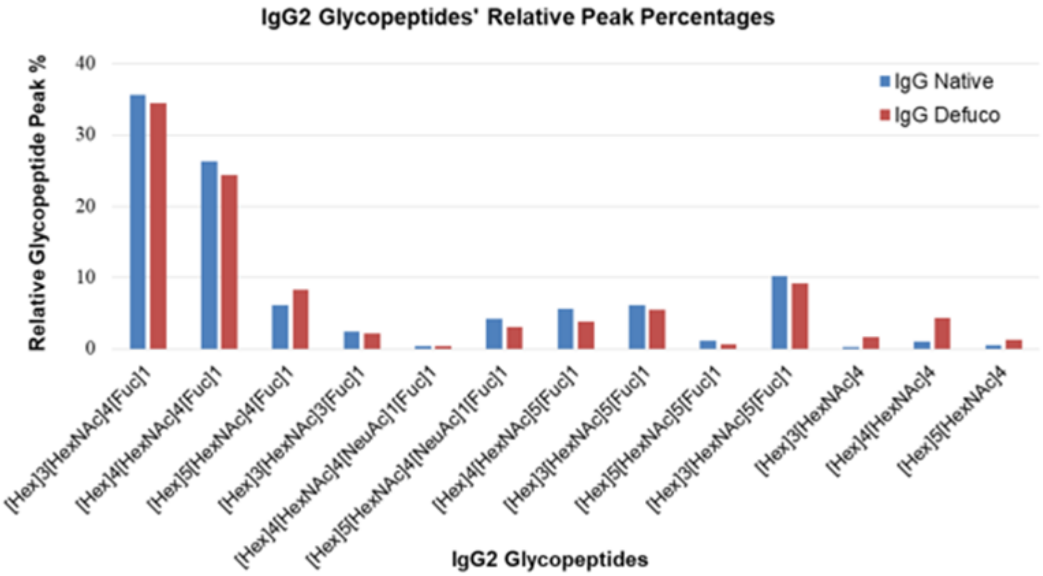


Figure 5. Relative peak percentages calculated for IgG2 glycopeptides generated from IgG native and IgG defuco tryptic digests.

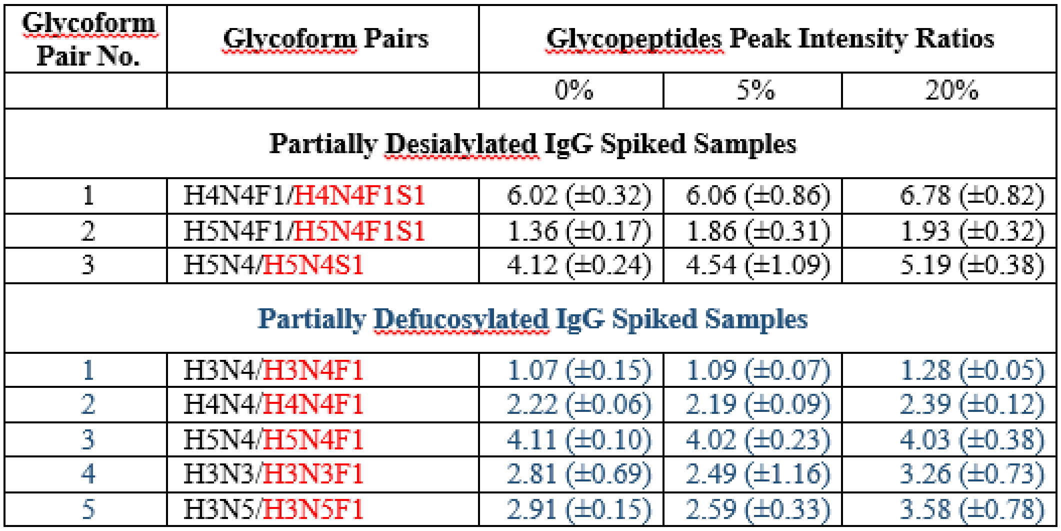
As shown in this figure, among ten fucosylated glycopeptides observed in the IgG defuco tryptic digest sample, almost all of the glycopeptides (except [Hex]5[HexNAc]4[Fuc]1) showed decreased relative glycopeptide peak percentage compared to that of the native IgG tryptic digest, as expected. The observed degree of defucosylation ranged in between (0.5 – 46) % at individual glycopeptide level. In addition, the corresponding defucosylated glycopeptides, the last three glycopeptides shown in Figure 5, showed increased relative glycopeptide peak percentages as compared to the control sample.

Based on the results obtained for IgG desialylated and IgG defucosylated tryptic digested samples, the sample preparation methods proved to be good enough to generate partially desialylated and partially defucosylated IgG glycopeptides for generating glycosylation altered samples as described next.

4.3.4 Identifying the Best Mixing Ratio of IgG Native and IgG Desialo/Defuco Digest

The final goal of this study was to generate large sets of IgG glycopeptides' LC-MS data to mimic samples of a healthy state (IgG native tryptic digest) and a disease state (a mixture of IgG native and IgG desialo/defuco tryptic digest). Therefore, after optimizing protocols to obtain partially desialylated or partially defucosylated IgG tryptic digested samples, the next step was to identify the amount of partially deglycosylated IgG that should be mixed with the native IgG to introduce slight changes to the glycosylation profile, mimicking a disease state sample. Therefore, a preliminary study was performed by spiking 0%, 5%, and 20% of partially desialylated or partially defucosylated IgG tryptic digest into corresponding IgG native tryptic digest samples, separately, while maintaining the final concentration of all the samples at a fixed concentration of 0.5 µg/µL. Table 2. represents the glycopeptide peak intensity ratios calculated for multiple IgG2 glycopeptide peak pairs; three and five glycopeptide pairs were selected for partially desialylated and partially defucosylated IgG spiked samples, respectively, to find out the best possible mixing ratio between the native and the partially deglycosylated IgG samples.

Table 2. IgG2 glycopeptide peak pair ratios calculated for 0%, 5%, and 20% of partially desialylated or partially defucosylated IgG2 tryptic digest spiked samples.



Glycoform Pair No.	**Glycoform Pairs**	**Glycopeptides Peak Intensity Ratios**		
		0%	5%	20%
	Partially Desialylated IgG Spiked Samples			
1	H4N4F1/H4N4F1S1	6.02 (±0.32)	6.06 (±0.86)	6.78 (±0.82)
2	H5N4F1/H5N4F1S1	1.36 (±0.17)	1.86 (±0.31)	1.93 (±0.32)
3	H5N4/H5N4S1	4.12 (±0.24)	4.54 (±1.09)	5.19 (±0.38)
	Partially Defucosylated IgG Spiked Samples			
1	H3N4/H3N4F1	1.07 (±0.15)	1.09 (±0.07)	1.28 (±0.05)
2	H4N4/H4N4F1	2.22 (±0.06)	2.19 (±0.09)	2.39 (±0.12)
3	H5N4/H5N4F1	4.11 (±0.10)	4.02 (±0.23)	4.03 (±0.38)
4	H3N3/H3N3F1	2.81 (±0.69)	2.49 (±1.16)	3.26 (±0.73)
5	H3N5/H3N5F1	2.91 (±0.15)	2.59 (±0.33)	3.58 (±0.78)

Each of these glycopeptide pairs contains a sialylated/fucosylated glycopeptide along with the corresponding desialylated/defucosylated glycopeptide. For an example, H4N4F1S1 and H4N4F1 represent a sialylated glycopeptide and its corresponding desialylated glycopeptide, respectively. The peak ratios for each glycopeptide pair were calculated by dividing the raw abundance recorded for the deglycosylated glycopeptide from that of the corresponding glycosylated glycopeptide in the same LC-MS run. Figure 6A and 6B illustrate the distribution of calculated IgG2 glycopeptides' peak intensity ratios of triplicate sample runs across 0%, 5%, and 20% desialylated (A) or defucosylated (B) IgG spiked samples.

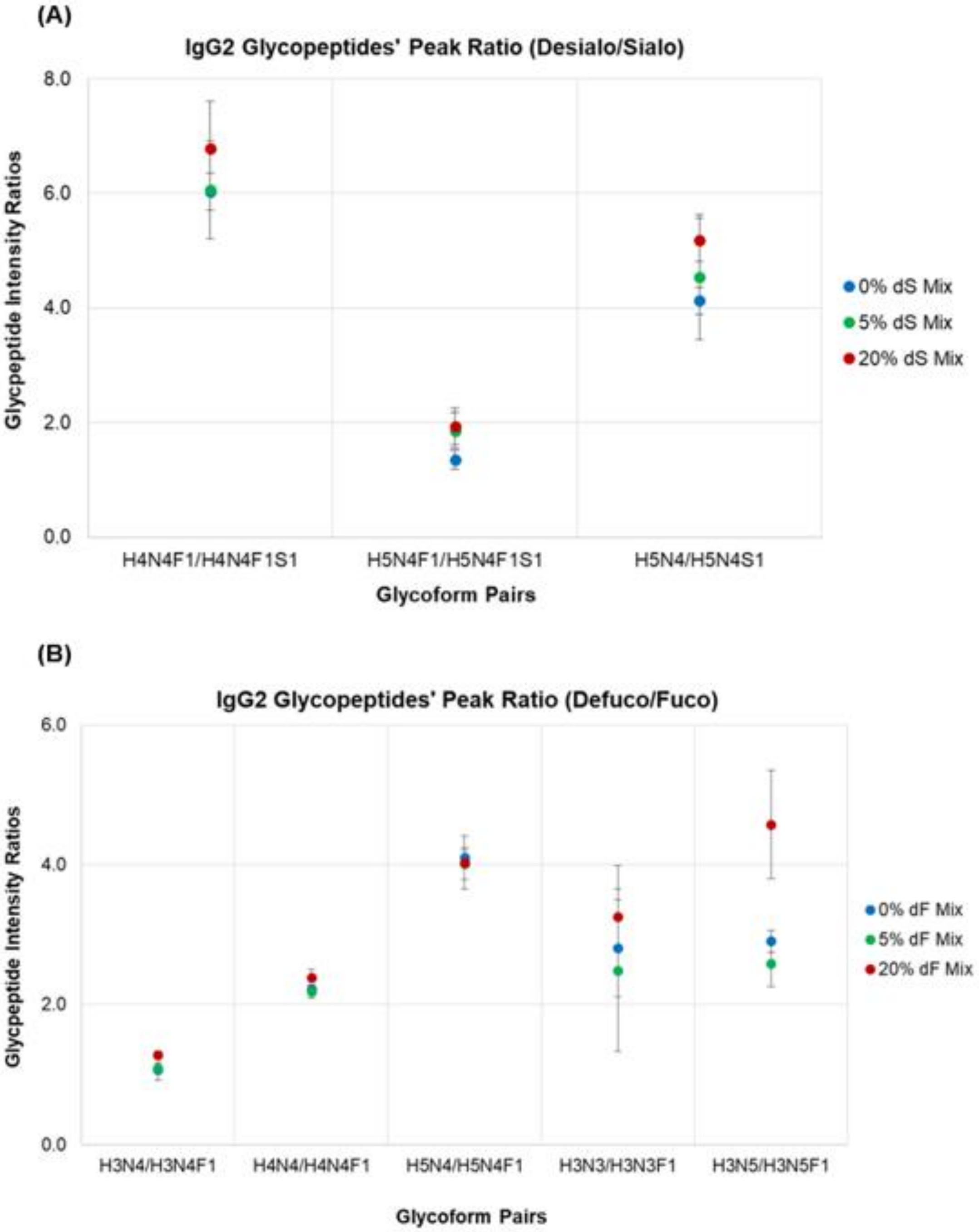


Figure 6. IgG2 glycopeptides' peak ratios calculated for multiple glycopeptide peak pairs identified in partially desialylated **(A)** or partially defucosylated **(B)** IgG spiked samples. 0% dS/dF mix sample represents native IgG tryptic digest at 0.5 µg/µL final concentration, while 5% and 20% dS/dF mix samples represent native IgG tryptic digest samples spiked with 5% and 20% of partially desialylated (dS) or partially defucosylated (dF) IgG tryptic digest at a fixed final concentration of 0.5 µg/µL. Error bars represent standard deviations (SDs) calculated for triplicate sample runs.

Based on the results, compared to the 0% desialo/defuco mixed samples (controls), glycopeptide peak pair ratios observed for 5% desialo/defuco spiked samples were very subtle and likely too similar to tell apart. By contrast, the 20% desialo/defuco mixed samples also showed subtle differences in glycopeptide peak ratios across many of the individual glycoform pairs, but their differences were somewhat more pronounced, so they would likely produce a data set that would be challenging to classify, but not so challenging that the samples could not be discriminated at all. Therefore, the 20% of partially desialylated/defucosylated IgG spiked mix sample was selected to mimic a disease state sample with subtle glycosylation differences.

4.3.5 Quantification of Large Sets of LC-MS Glycopeptides Data of IgG Native and IgG Samples with Altered Sialylation

To optimize The Aristotle Classifier, the classification algorithm that uses the whole glycomic profile to classify samples, a set of optimized data were necessary; thus, we sought to chemically generate large sets of LC-MS data for IgG glycopeptides to mimic samples of a healthy state and a disease state. Of these samples, healthy state sample composed of a tryptic digest of IgG with a native glycosylation profile while the disease state sample contains IgG native tryptic digest purposely adulterated with 20% of partially desialylated or partially defucosylated IgG tryptic digest. Furthermore, to effectively use these two sets of data to optimize the Aristotle Classifier, they needed to be different, but not so different where any of the glycopeptide peak heights of these samples can tell them apart.

As described in the experimental section 4.2.5, as the first set of samples, IgG sialo native and IgG desialo 20% mixed samples were prepared; the samples were slightly different in terms of their sialylation profile; these samples were used to generate group 1 and group 2 data. Group 1 includes 21 samples, each containing IgG native tryptic digest, while group 2 samples contains

80% of native IgG tryptic digest mixed with 20% of partially desialylated IgG sample. Both IgG1 and IgG2 glycopeptides were detected in LC-MS runs acquired for group 1 and group 2 samples (21 for each group); these data were separately analyzed and quantified; see Appendix C, Table 1 and Table 2 for percent of different glycopeptides obtained for both IgG1 and IgG2 glycopeptides across 42 samples. The quantitation results obtained for IgG1 and IgG2 glycopeptides are represented in Figure 7A and 7B, respectively. Figure 7A represents fifteen different IgG1 glycopeptides' percentages calculated for group 1 and group 2 samples; the data showed that the samples of group 1 and group 2 are not distinguishable based on any single glycopeptide in the group. Similarly, Figure 7B represents thirteen different IgG2 glycopeptides quantified for the same group 1 and group 2 samples; again, any of the quantified glycopeptides were not useful in distinguishing samples of group 1 and group 2. These data clearly showed that the glycosylation differences among two groups of samples are very subtle; thus, these data sets represent a challenging classification problem that can be used to test different classification tools. Therefore, we submitted these data to the existing standard approach, PCA, and also to the Aristotle Classifier, the classifier that our group has developed, with the aim of comparing the classification powers of each method.

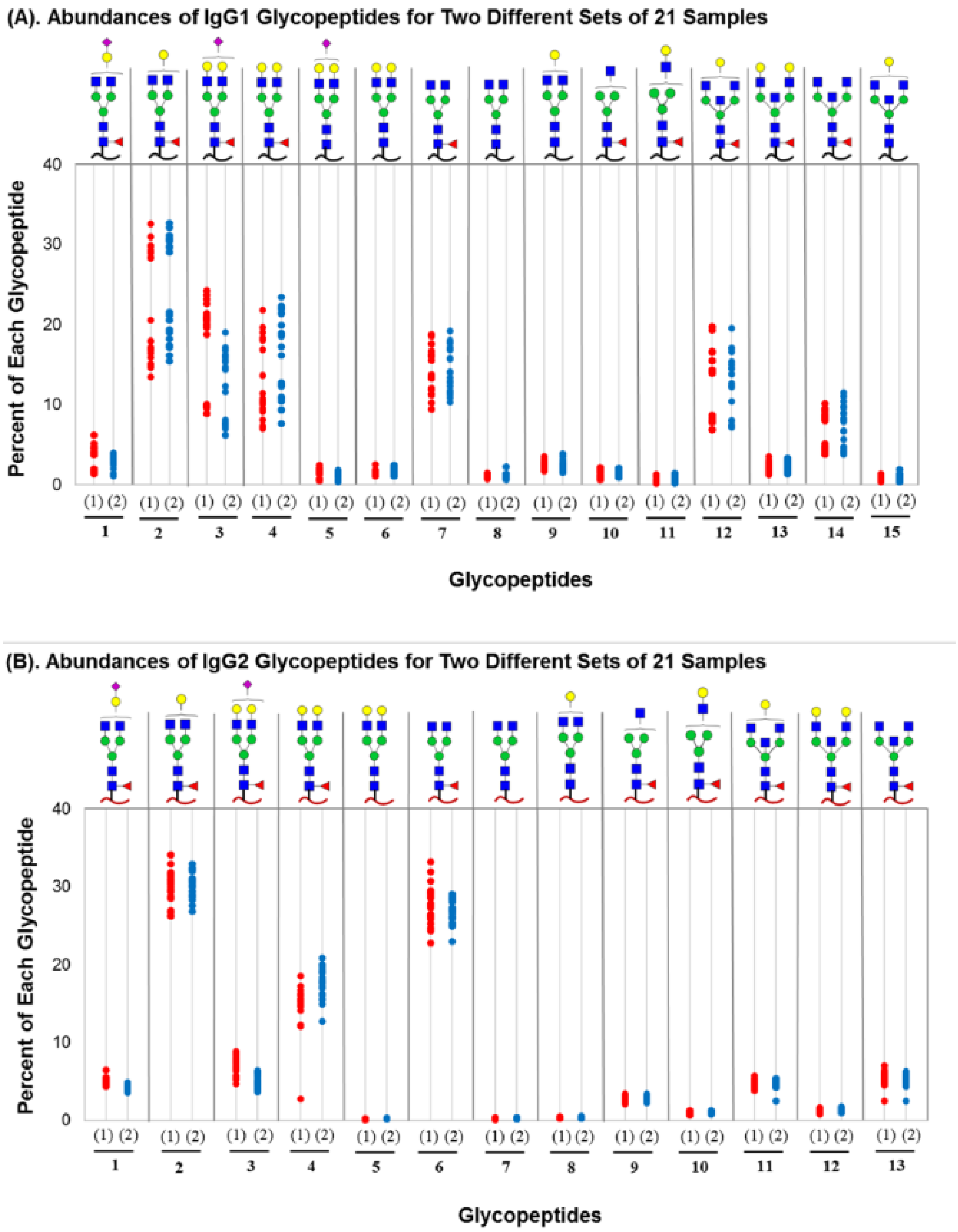


Figure 7. Comparison of the abundance of 15 different IgG1 glycopeptides **(A)** and 13 different IgG2 glycopeptides **(B)** quantified in two groups of IgG samples. Group 1 includes 21 native IgG samples; group 2 includes 21 IgG samples with slightly altered sialylation (20% desialo mixed). The percent of each different IgG glycopeptides is plotted for both group 1 and group 2; group 1 data are shown in the left (red dots) for each different IgG glycopeptide, while group 2 data are shown on the right (blue dots). Figure 7B is reproduced from reference 19 with permission from the *American Chemical Society*.

Figure 8A and 8B represent the PCA plot of the IgG1 glycopeptides' data originating from group 1 and group 2 sample sets (A) and the output plot of the same data from the Aristotle Classifier (B). Similarly, Figure 8C and 8D represent output plots of PCA and the Aristotle Classifier for the IgG2 glycopeptides' data.

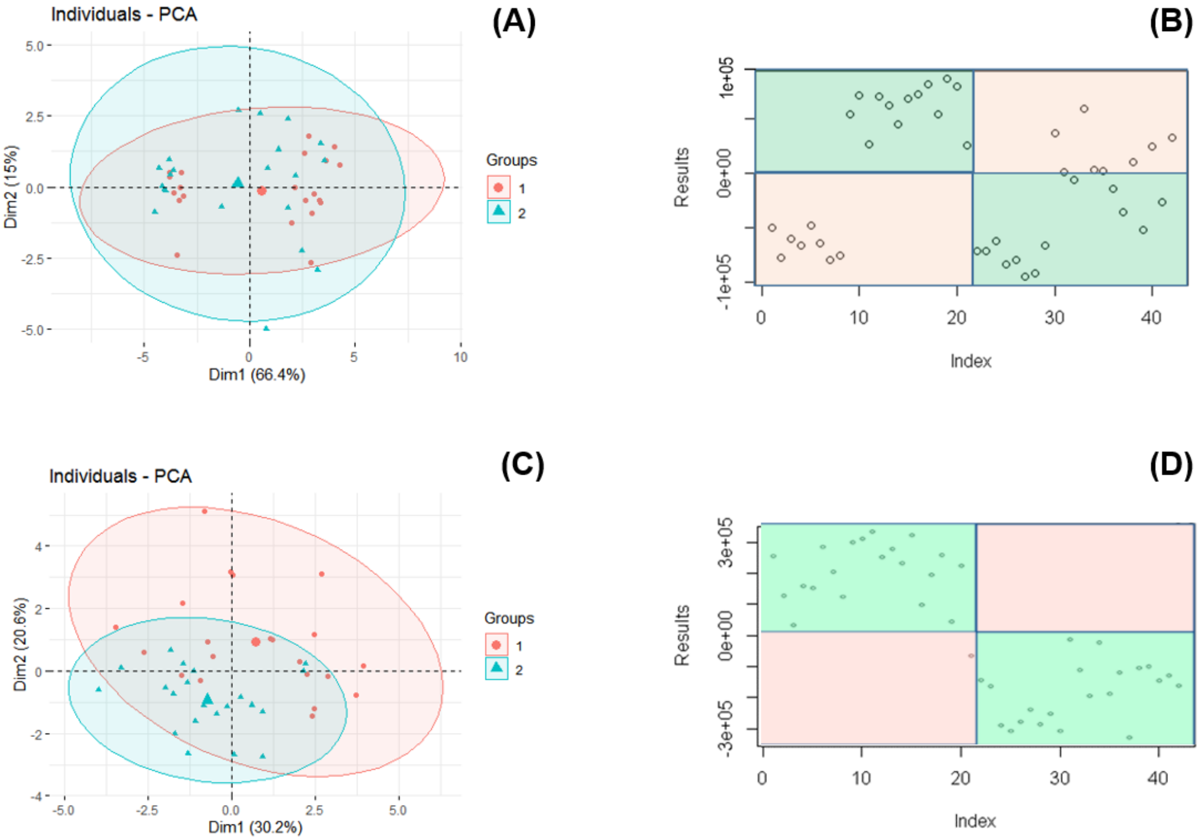


Figure 8. Comparison of two different data sets (group 1 and group 2) by PCA and the Aristotle Classifier. **(A)** PCA classification of a total of 42 IgG samples based on IgG1 glycopeptodes' quantitation data; samples were from group 1 and group 2; each group with 21 samples. **(B)** Classification of the same set of 42 IgG samples of panel A by using the Aristotle Classifier; the Aristotle Classifier accurately assigned 26/42 samples into their correct groups. Results of PCA **(C)** and the Aristotle Classifier **(D)** classification of the same 42 IgG samples based on IgG2 glycopeptides' quantitation data. PCA correctly classified 30/42 samples, while the Aristotle Classifier assigned 41 of the 42 samples accurately. The correctly classified samples are in green area of the Aristotle Classifier plots. Figures 8C and 8D are reproduced from reference 19 with permission from the *American Chemical Society*.

Based on the outcome of the PCA and the Aristotle Classsifier, the latter proved to be better than the PCA for discriminating generated large sets of glycomic data; for IgG1 glycopeptides' samples; as shown in Figure 8A, PCA showed substantial overalap between the samples of both healthy and disease states while showing that the variability of group 1 and group 2 samples are not related to the glycosylation differences between two sample groups. Compared to the poor classification performed by PCA for classifying samples of group1 and group 2, based on IgG1 glycopeptides' data, the Aristotle Classifier performed better; see Figure 8B; it accurately classified 26/42 samples (62% accuracy) into their accurate groups based on their glycosylation differences. Furthermore, for IgG2 glycopeptides' samples, 98% of the data were classified correctly by the Aristotle Classifier (D), in contrast to the PCA (C), where a considerable overlapping of the samples from two groups were observed, resulting only 30 of the 42 samples to be correctly classified into accurate groups. These results clearly show that the Aristotle Classifier can handle challenging glycomics data while showing its capacity to distinguish samples based on their glycosylation differences.

4.3.6 Quantification of Large Sets of LC-MS Glycopeptides Data of IgG Native and IgG Samples with Altered Fucosylation

Not only the degree of sialylation, but also the degree of fucosylation of the IgG samples were slightly altered during this study to generate two groups of IgG samples, IgG fuco native (group 3) and IgG defuco 20% mixed samples (group 4). Among the two groups; group 3 composed of 19 IgG samples with native IgG glycosylation profile and the group 4 composed of 19 IgG samples, those fucosylation profiles that were slightly altered purposely by mixing 80% of the native IgG tryptic digest with a 20% of partially defucosylated IgG tryptic digest; see sections 4.2.5 to 4.2.7 for detailed sample preparation and data analysis performed for the group

3 and group 4 samples. During the LC-MS runs of these samples, the retention time of the IgG1 glycopeptides varied significantly between multiple LC-MS runs; therefore, for this data set with 38 samples (19 samples per each group), only IgG2 glycopeptides were considered for the quantitation.

Appendix C, Table 3 includes IgG2 glycopeptide peak ratio percentages obtained for group 3 (IgG fuco native sample) and group 4 (IgG defuco 20% mixed sample); 19 samples were acquired for each group in two different days. The quantitation results obtained for thirteen different IgG2 glycopeptides generated for group 3 and group 4 samples are shown in Figure 9A; this figure clearly showed that the classification of group 3 and group 4 samples is difficult based on any of the individual glycopeptides in the samples.

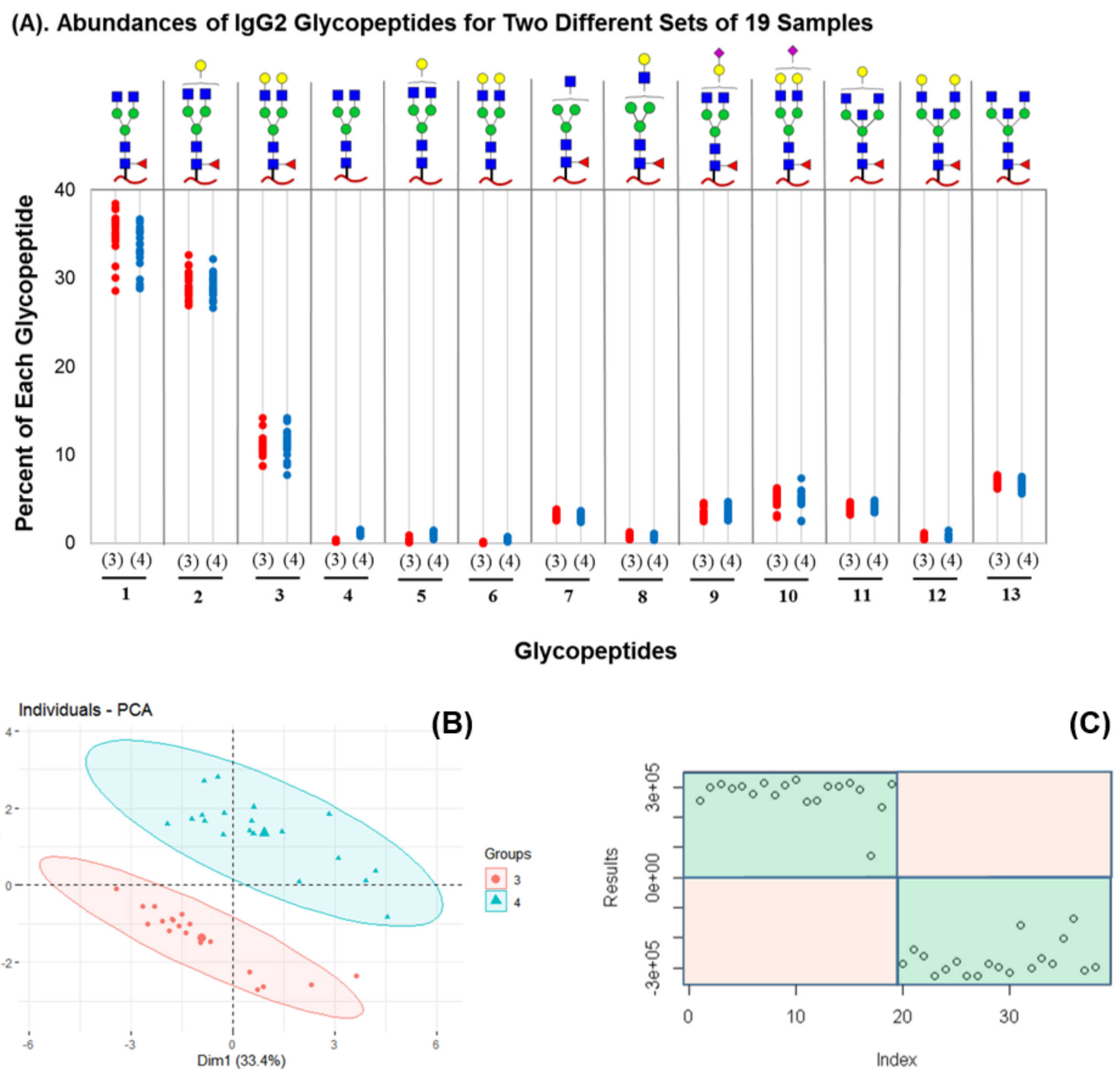


Figure 9. Comparison of the abundance of 13 different IgG2 glycopeptides relative percentages of group 3 (native IgG) and group 4 (20% defuco mixed) IgG samples **(A)**. Group 3 data are shown in the left (red dots) while group 4 data are shown on the right (blue dots). PCA plot generated for 19 native IgG samples (group 3) and 19 IgG samples with slightly altered fucosylation (group 4) **(B)**. Classification of the same set of 38 IgG samples based on the IgG2 glycopeptides data by using the Aristotle Classifier **(C)**.

Figure 9B represents the classification of 38 IgG samples by using PCA; in this example, PCA clearly separated IgG samples of group 3 and group 4 into their accurate groups. This indicates that the introduced fucosylation changes (20% mix) to the IgG native glycosylation profile contributed significantly, which in turn resulted better separation of the group 3 and group 4 samples by PCA. The classification of the same 38 IgG samples with the Aristotle Classifier is shown in Figure 9C; similar to PCA, it also separated all the samples accurately into their groups

Overall, the generated IgG data, mimicking two different biological states, were useful to challenge both PCA and the newly developed Aristotle Classifier. The classification results clearly showed that the Aristotle Classifier, which uses supervised classification on the whole glycomic profile, performed far better than PCA and thus, it can be applied to differentiate challenging clinical glycomics samples.

4.4 Conclusion

Multiple data sets of glycomics data were developed to study the merits of a new classifier, the Aristotle Classifier, a supervised machine classification algorithm, which uses the whole glycomic profile to classify glycomics samples. We used IgG glycoprotein to chemically generate two challenging glycomics sample groups mimicking two different biological states as healthy and disease; of these samples, healthy state represented samples with a native IgG glycosylation profile, while the disease state represented samples with a slightly altered IgG glycosylation profile. We optimized both the sample preparation and the LC-MS conditions to effectively generate large sets of IgG glycopeptides data of two sample groups; the resulting glycopeptides' LC-MS data were quantified; however, those samples could not be classified into native or disease state by considering any of the individual glycopeptides of the samples. Therefore, these data proved to be challenging; thus, they were used to challenge the classification ability of a standard approach: PCA and the newly developed algorithm: the

Aristotle Classifier. The results showed that the classification ability of the Aristotle Classifier outperformed the PCA, while successfully classifying many of the glycomics samples into their accurate groups.

4.5 Acknowledgements

These studies were funded by R01GM103547 and R35GM130354 to Heather Desaire. We are truly thankful for their generous support.

4.6 References

1. Ruhaak, L. R.; Barkauskas, D. A.; Torres, J.; Cooke, C. L.; Wu, L. D.; Stroble, C.; Ozcan, S.; Williams, C. C.; Camorlinga, M.; Rocke, D. M.; Lebrilla, C. B.; Solnick, J. V., *EuPA open proteomics* **2015,** *6*, 1-9.
2. Shih, H. C.; Chang, M. C.; Chen, C. H.; Tsai, I. L.; Wang, S. Y.; Kuo, Y. P.; Chen, C. H.; Chang, Y. T., *Clinical proteomics* **2019,** *16*, 1.
3. Ruhaak, L. R.; Kim, K.; Stroble, C.; Taylor, S. L.; Hong, Q.; Miyamoto, S.; Lebrilla, C. B.; Leiserowitz, G., *Journal of proteome research* **2016,** *15* (3), 1002-10.
4. Jia, X.; Chen, J.; Sun, S.; Yang, W.; Yang, S.; Shah, P.; Hoti, N.; Veltri, B.; Zhang, H., *Proteomics* **2016,** *16* (23), 2989-2996.
5. Tsai, T. H.; Wang, M.; Di Poto, C.; Hu, Y.; Zhou, S.; Zhao, Y.; Varghese, R. S.; Luo, Y.; Tadesse, M. G.; Ziada, D. H.; Desai, C. S.; Shetty, K.; Mechref, Y.; Ressom, H. W., *Journal of proteome research* **2014,** *13* (11), 4859-68.
6. Zhang, X.; Wang, Y.; Qian, Y.; Wu, X.; Zhang, Z.; Liu, X.; Zhao, R.; Zhou, L.; Ruan, Y.; Xu, J.; Liu, H.; Ren, S.; Xu, C.; Gu, J., *PLoS One* **2014,** *9* (2), e87978.
7. Ruhaak, L. R.; Taylor, S. L.; Stroble, C.; Nguyen, U. T.; Parker, E. A.; Song, T.; Lebrilla, C. B.; Rom, W. N.; Pass, H.; Kim, K.; Kelly, K.; Miyamoto, S., *Journal of proteome research* **2015,** *14* (11), 4538-49.
8. Liu, S.; Cheng, L.; Fu, Y.; Liu, B. F.; Liu, X., *J Proteomics* **2018,** *181*, 225-237.
9. Russell, A. C.; Simurina, M.; Garcia, M. T.; Novokmet, M.; Wang, Y.; Rudan, I.; Campbell, H.; Lauc, G.; Thomas, M. G.; Wang, W., *Glycobiology* **2017,** *27* (5), 501-510.

10. Menni, C.; Gudelj, I.; Macdonald-Dunlop, E.; Mangino, M.; Zierer, J.; Besic, E.; Joshi, P. K.; Trbojevic-Akmacic, I.; Chowienczyk, P. J.; Spector, T. D.; Wilson, J. F.; Lauc, G.; Valdes, A. M., *Circulation research* **2018,** *122* (11), 1555-1564.

11. Argade, S.; Chen, T.; Shaw, T.; Berecz, Z.; Shi, W.; Choudhury, B.; Parsons, C. L.; Sur, R. L., *Urolithiasis* **2015,** *43* (4), 303-12.

12. Vivekanandan-Giri, A.; Slocum, J. L.; Buller, C. L.; Basrur, V.; Ju, W.; Pop-Busui, R.; Lubman, D. M.; Kretzler, M.; Pennathur, S., *International journal of proteomics* **2011,** *2011*, 214715.

13. Lundstrom, S. L.; Yang, H.; Lyutvinskiy, Y.; Rutishauser, D.; Herukka, S. K.; Soininen, H.; Zubarev, R. A., *Journal of Alzheimer's disease : JAD* **2014,** *38* (3), 567-79.

14. Varki, A., *Glycobiology* **1993,** *3* (2), 97-130.

15. Dwek, R. A., *Chemical reviews* **1996,** *96* (2), 683-720.

16. Zhang, Y.; Go, E. P.; Desaire, H., *Anal Chem* **2008,** *80* (9), 3144-58.

17. Tan, Z.; Yin, H.; Nie, S.; Lin, Z.; Zhu, J.; Ruffin, M. T.; Anderson, M. A.; Simeone, D. M.; Lubman, D. M., *Journal of proteome research* **2015,** *14* (4), 1968-78.

18. Stockmann, H.; Coss, K. P.; Rubio-Gozalbo, M. E.; Knerr, I.; Fitzgibbon, M.; Maratha, A.; Wilson, J.; Rudd, P.; Treacy, E. P., *JIMD reports* **2016,** *27*, 47-53.

19. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., *Anal Chem* **2019,** *91* (17), 11070-11077.

20. Vidarsson, G.; Dekkers, G.; Rispens, T., *Frontiers in immunology* **2014,** *5*, 520.

21. Li, T.; DiLillo, D. J.; Bournazos, S.; Giddens, J. P.; Ravetch, J. V.; Wang, L. X., *Proc Natl Acad Sci U S A* **2017,** *114* (13), 3485-3490.

22. Castilho, A.; Gruber, C.; Thader, A.; Oostenbrink, C.; Pechlaner, M.; Steinkellner, H.; Altmann, F., *mAbs* **2015,** *7* (5), 863-70.

23. Reusch, D.; Haberger, M.; Maier, B.; Maier, M.; Kloseck, R.; Zimmermann, B.; Hook, M.; Szabo, Z.; Tep, S.; Wegstein, J.; Alt, N.; Bulau, P.; Wuhrer, M., *mAbs* **2015,** *7* (1), 167-79.

24. Goetze, A. M.; Liu, Y. D.; Zhang, Z.; Shah, B.; Lee, E.; Bondarenko, P. V.; Flynn, G. C., *Glycobiology* **2011,** *21* (7), 949-59.

25. Kyselova, Z.; Mechref, Y.; Al Bataineh, M. M.; Dobrolecki, L. E.; Hickey, R. J.; Vinson, J.; Sweeney, C. J.; Novotny, M. V., *Journal of proteome research* **2007,** *6* (5), 1822-32.

26. Hu, W.; Su, X.; Zhu, Z.; Go, E. P.; Desaire, H., *Analytical and bioanalytical chemistry* **2017,** *409* (2), 561-570.

4.7 Appendix C

Table 1. IgG1 glycopeptide peak ratio percentages obtained for group 1(IgG sialo native sample) and group 2 (IgG desialo 20% mixed sample), 21 samples were acquired for each group in two different days.

Gly. Peptides	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Group	Data Set
V1	2.00	29.30	9.66	13.63	0.76	2.54	18.76	1.45	2.87	2.08	1.23	8.36	1.69	5.14	0.54	1	first
V2	1.45	28.97	8.92	21.81	0.53	1.45	17.61	1.22	2.47	1.70	1.22	6.89	1.51	3.76	0.49	1	first
V3	1.60	29.76	9.93	19.70	0.64	1.86	16.05	1.19	2.56	1.75	1.29	7.98	1.42	3.80	0.48	1	first
V4	1.80	28.28	10.10	19.02	0.59	1.86	16.28	1.30	2.39	1.73	1.11	8.76	1.55	4.73	0.50	1	first
V5	1.72	30.99	10.01	18.27	0.66	1.80	16.27	1.38	2.45	1.58	1.08	7.82	1.47	4.09	0.42	1	first
V6	1.64	29.90	9.82	18.36	0.71	1.78	16.67	1.17	2.47	1.85	1.19	7.82	1.70	4.35	0.55	1	first
V13	1.79	28.47	9.73	18.08	0.68	1.58	17.65	1.04	2.44	2.15	1.25	8.02	1.73	4.74	0.65	1	Second
V14	1.41	32.54	8.95	16.87	0.54	1.58	18.56	1.11	2.26	1.85	1.15	6.94	1.28	4.41	0.55	1	Second
V15	4.59	17.03	20.86	10.11	2.49	1.39	10.22	0.87	1.83	0.68	0.49	16.49	2.51	9.15	1.27	1	Second
V16	4.73	17.95	20.03	9.22	1.82	1.43	11.24	1.19	2.85	0.82	0.42	16.73	2.27	7.98	1.32	1	Second
V17	5.17	20.60	18.78	10.38	1.62	1.12	11.32	0.95	1.87	1.23	0.61	14.10	2.68	8.19	1.39	1	Second
V18	4.11	17.97	19.63	7.42	2.01	1.70	16.06	1.26	2.71	1.08	0.46	14.33	1.80	8.50	0.96	1	Second
V19	3.83	13.47	24.27	7.19	1.72	1.48	11.32	0.78	2.12	0.87	0.22	19.73	3.50	8.64	0.86	1	Second
V20	3.96	17.16	20.52	10.37	2.05	1.12	13.46	1.18	1.82	1.22	0.74	14.30	2.78	8.07	1.26	1	Second
V21	4.11	16.79	20.99	9.42	1.95	1.34	13.30	1.15	2.69	0.94	0.47	14.31	2.83	8.55	1.14	1	Second
V22	4.09	14.61	23.73	9.50	1.37	1.52	13.29	1.11	2.57	0.76	0.39	14.12	2.10	10.15	0.69	1	Second
V23	6.22	14.77	22.59	10.79	1.74	1.47	11.90	1.16	2.70	0.69	0.34	14.26	2.05	8.29	1.03	1	Second
V24	3.72	15.89	23.21	8.12	1.49	1.71	12.02	0.92	3.16	1.01	0.37	15.54	2.45	9.49	0.91	1	Second
V25	4.58	16.48	21.41	6.99	1.92	1.73	13.74	1.49	3.53	0.83	0.28	15.48	2.08	8.53	0.93	1	Second
V26	4.11	15.13	20.86	10.56	1.91	1.60	15.54	0.95	2.32	0.98	0.51	13.97	1.74	8.86	0.98	1	Second
V27	3.76	15.10	20.70	11.39	1.36	1.33	9.42	1.28	1.70	0.88	0.23	19.35	3.06	9.37	1.05	1	Second
V7	1.36	30.63	6.22	19.91	0.59	2.31	17.77	1.26	2.45	2.09	1.40	7.45	1.42	4.72	0.42	2	first
V8	1.38	31.13	7.00	22.37	0.41	1.68	16.89	1.39	2.34	1.37	1.22	7.19	1.46	3.83	0.35	2	first
V9	1.15	30.40	8.09	22.06	0.52	1.75	16.89	1.14	2.19	1.48	1.09	7.39	1.47	3.96	0.43	2	first
V10	1.30	29.65	7.31	21.45	0.48	1.81	18.03	1.32	2.26	1.72	1.19	7.24	1.55	4.15	0.54	2	first
V11	1.29	29.82	7.73	21.87	0.50	1.67	15.72	1.22	2.17	1.87	1.17	8.02	1.75	4.65	0.54	2	first
V12	1.21	32.69	7.59	23.44	0.54	1.71	11.61	1.17	2.26	2.03	1.34	8.01	1.59	4.34	0.48	2	first
V28	1.21	30.75	6.23	21.40	0.58	1.65	17.68	1.06	2.34	2.06	1.45	7.36	1.65	4.07	0.52	2	Second
V29	1.11	30.60	7.66	18.92	0.49	1.90	19.23	1.17	2.18	1.83	1.25	7.47	1.63	4.12	0.46	2	Second
V30	3.59	17.11	15.36	9.32	1.48	2.11	13.40	1.41	3.04	0.93	0.49	16.60	2.47	11.52	1.18	2	Second
V31	3.35	17.41	15.89	10.98	1.08	1.14	12.75	0.97	1.99	0.89	0.48	17.03	3.07	11.10	1.88	2	Second
V32	2.74	32.13	14.73	9.37	1.45	1.07	10.30	0.97	1.51	1.31	0.91	12.35	2.15	8.26	0.74	2	Second
V33	3.90	16.18	19.07	10.83	1.61	1.35	11.54	0.90	2.18	1.05	0.21	17.04	3.10	9.78	1.26	2	Second
V34	2.85	20.59	16.12	12.36	1.17	1.57	11.82	1.39	2.72	0.89	0.50	14.57	2.88	9.59	0.97	2	Second
V35	3.66	19.41	16.80	10.47	1.19	2.08	13.99	0.82	1.90	0.92	0.55	14.99	2.86	9.02	1.36	2	Second
V36	3.27	19.07	15.51	16.17	1.55	1.30	10.85	0.91	2.42	1.43	1.13	14.53	2.60	7.93	1.33	2	Second
V37	2.44	20.54	14.48	12.51	1.02	1.44	14.07	0.87	2.45	1.50	0.70	15.39	2.89	8.94	0.76	2	Second
V38	4.01	21.16	17.13	12.73	1.66	1.12	12.93	0.65	1.85	1.36	0.74	12.67	2.54	8.18	1.26	2	Second
V39	2.00	29.12	12.29	17.22	1.05	1.17	11.25	0.85	1.54	1.10	0.87	12.25	1.88	6.70	0.72	2	Second
V40	3.22	18.22	14.36	12.30	1.81	2.49	15.84	2.24	3.54	0.99	0.59	13.80	2.93	6.79	0.89	2	Second
V41	3.08	21.54	11.60	18.57	1.04	1.90	17.20	1.16	2.61	1.26	1.12	10.39	2.06	5.72	0.74	2	Second
V42	4.02	15.42	16.24	7.66	1.18	2.48	12.30	0.92	3.90	1.16	0.30	19.58	3.31	10.41	1.11	2	Second

Table 2. IgG2 glycopeptide peak ratio percentages obtained for group 1(IgG sialo native sample) and group 2 (IgG desialo 20% mixed sample); 21 samples were acquired for each group in two different days.

Gly. Peptides	1	2	3	4	5	6	7	8	9	10	11	12	13	Group	Data set
V1	5.27	32.90	6.33	15.60	0.10	27.60	0.11	0.25	2.50	0.84	4.73	1.36	2.50	1	first
V2	4.60	34.10	4.68	12.10	0.11	28.60	0.19	0.29	3.15	0.72	4.82	1.02	5.62	1	first
V3	5.46	31.40	5.29	16.00	0.14	26.40	0.25	0.34	2.86	1.09	4.84	1.35	4.52	1	first
V4	5.26	31.60	5.63	17.20	0.12	24.70	0.20	0.26	2.41	1.06	5.22	1.46	4.88	1	first
V5	4.88	31.00	6.96	18.50	0.15	22.80	0.19	0.31	2.56	1.14	5.35	1.25	4.97	1	first
V6	6.48	34.00	6.73	2.72	0.15	31.90	0.24	0.31	3.09	1.09	5.69	1.47	6.13	1	first
V13	5.16	31.80	5.35	12.20	0.09	29.20	0.25	0.23	3.15	0.93	5.00	1.12	5.62	1	second
V14	4.95	30.60	5.69	16.70	0.13	26.30	0.25	0.31	2.75	1.00	5.16	1.35	4.75	1	second
V15	5.29	29.60	6.75	14.10	0.11	28.90	0.20	0.46	3.00	0.72	4.96	1.10	4.82	1	second
V16	5.21	26.70	7.58	12.30	0.14	33.20	0.29	0.38	2.80	0.95	4.23	0.91	5.28	1	second
V17	5.55	28.50	8.39	14.70	0.14	27.30	0.23	0.37	2.94	1.17	4.48	1.28	4.84	1	second
V18	5.03	29.32	8.83	14.71	0.17	26.08	0.27	0.35	2.36	1.20	4.48	1.18	6.00	1	second
V19	5.03	30.91	8.80	15.22	0.16	24.38	0.29	0.37	2.68	0.90	4.78	1.22	5.20	1	second
V20	5.02	28.77	7.25	15.28	0.15	29.21	0.25	0.41	2.07	0.86	3.89	1.33	5.46	1	second
V21	5.05	26.94	8.44	16.18	0.13	27.81	0.21	0.42	3.13	0.84	3.88	1.09	5.83	1	second
V22	4.71	29.39	7.42	16.26	0.18	25.28	0.32	0.29	3.36	0.79	4.19	1.21	6.49	1	second
V23	4.72	26.27	7.68	15.44	0.14	29.23	0.31	0.46	2.70	0.97	4.23	0.79	6.99	1	second
V24	4.33	26.22	7.94	14.89	0.08	30.75	0.24	0.34	2.66	0.97	4.41	1.28	5.84	1	second
V25	4.89	29.95	8.05	15.55	0.25	25.82	0.26	0.36	2.35	1.00	4.49	1.64	5.39	1	second
V26	4.94	26.21	7.97	15.44	0.17	29.48	0.31	0.42	2.41	0.74	4.91	0.91	6.06	1	second
V27	4.86	30.35	6.75	15.55	0.18	26.45	0.36	0.40	2.67	1.31	4.50	1.34	5.24	1	second
V7	3.83	32.00	5.01	19.00	0.19	26.90	0.14	0.27	2.80	1.19	4.77	1.45	2.49	2	first
V8	4.33	32.20	4.70	20.00	0.15	23.00	0.21	0.30	2.53	1.15	4.84	1.74	4.81	2	first
V9	3.82	30.60	4.68	20.80	0.16	25.10	0.22	0.32	2.71	1.14	4.67	1.37	4.37	2	first
V10	3.87	30.00	4.31	20.00	0.17	26.10	0.23	0.38	2.56	1.13	4.98	1.59	4.66	2	first
V11	4.04	30.80	4.49	17.80	0.19	26.60	0.27	0.25	2.83	1.05	5.33	1.48	4.95	2	first
V12	4.27	32.40	4.00	19.00	0.15	24.90	0.22	0.31	2.85	1.22	4.98	1.29	4.46	2	first
V28	4.23	32.96	4.02	16.26	0.18	26.10	0.24	0.33	2.92	1.16	4.96	1.40	5.16	2	second
V29	4.05	31.04	3.68	17.71	0.16	27.32	0.20	0.28	2.87	1.09	5.01	1.40	5.16	2	second
V30	3.58	31.00	5.27	17.25	0.16	27.10	0.29	0.39	2.59	1.13	4.73	1.35	5.14	2	second
V31	3.84	28.92	6.30	14.90	0.14	28.44	0.39	0.40	3.38	1.08	4.80	1.04	6.31	2	second
V32	3.97	32.38	5.61	12.71	0.28	28.78	0.32	0.35	3.19	0.87	4.55	1.06	5.87	2	second
V33	4.11	28.79	4.96	17.94	0.16	28.41	0.29	0.46	2.37	0.83	4.79	1.21	5.65	2	second
V34	4.64	30.70	5.91	15.59	0.17	27.09	0.30	0.50	2.87	0.94	4.68	1.05	5.49	2	second
V35	4.07	30.05	6.23	18.38	0.26	27.08	0.29	0.43	2.91	0.97	2.51	1.31	5.48	2	second
V36	4.29	26.80	6.37	19.25	0.17	26.95	0.28	0.41	3.05	1.11	4.59	1.33	5.38	2	second
V37	4.28	30.27	5.02	17.70	0.35	26.00	0.29	0.40	2.45	1.23	4.87	1.40	5.69	2	second
V38	4.16	28.31	6.24	15.49	0.24	29.05	0.36	0.30	3.16	0.91	4.23	1.22	6.28	2	second
V39	4.81	28.82	6.30	16.04	0.24	28.07	0.38	0.44	2.54	1.01	4.53	1.44	5.38	2	second
V40	4.40	29.46	4.89	19.07	0.13	25.92	0.34	0.38	2.38	1.06	5.31	1.23	5.42	2	second
V41	4.89	27.55	6.17	16.96	0.21	28.55	0.37	0.43	2.34	1.17	4.24	1.20	5.88	2	second
V42	3.71	29.38	6.24	19.61	0.16	25.34	0.33	0.41	2.34	1.24	4.41	1.15	5.65	2	second

Table 3. IgG2 glycopeptide peak ratio percentages obtained for group 3 (IgG fuco native sample) and group 4 (IgG defuco 20% mixed sample); 19 samples were acquired for each group in two different days.

Gly. Peptides	1	2	3	4	5	6	7	8	9	10	11	12	13	Group	Data set
V1	36.77	30.04	9.88	0.16	0.46	0.07	3.82	0.83	2.46	3.22	3.92	0.66	7.71	1	first
V2	35.65	30.74	10.93	0.28	0.28	0.07	3.46	0.80	3.66	2.96	4.28	0.76	6.16	1	first
V3	34.94	30.30	11.44	0.25	0.27	0.04	3.62	0.54	3.10	4.34	4.20	0.66	6.30	1	first
V4	34.21	31.48	10.56	0.14	0.27	0.10	3.50	0.57	3.11	4.34	4.67	0.77	6.26	1	first
V5	36.47	30.62	8.72	0.14	0.20	0.06	3.25	0.63	3.25	4.84	4.12	0.86	6.83	1	first
V11	36.23	28.02	11.38	0.26	0.41	0.08	3.34	0.71	3.23	5.05	3.25	0.70	7.31	1	second
V12	37.82	28.22	8.78	0.24	0.35	0.07	3.25	0.63	2.83	5.44	4.29	0.61	7.47	1	second
V13	36.10	28.99	10.42	0.32	0.26	0.11	3.57	0.43	3.29	4.91	4.23	0.52	6.84	1	second
V14	36.63	28.36	10.26	0.41	0.28	0.04	3.43	0.57	3.44	4.82	3.74	0.63	7.38	1	second
V15	38.46	27.54	10.61	0.20	0.20	0.06	3.24	0.47	3.15	4.82	3.96	0.47	6.80	1	second
V16	36.57	28.69	10.89	0.24	0.22	0.16	2.93	1.29	2.79	5.13	4.37	0.45	6.27	1	second
V17	38.02	26.89	11.24	0.40	0.24	0.12	3.19	0.61	3.46	4.79	3.65	0.55	6.84	1	second
V18	34.54	28.99	11.85	0.29	0.36	0.07	3.37	0.45	3.14	5.22	3.82	0.46	7.43	1	second
V19	33.63	29.74	11.53	0.27	0.20	0.10	3.27	0.65	3.21	5.84	3.68	0.59	7.27	1	second
V20	34.91	28.18	10.75	0.28	0.09	0.02	2.61	0.84	4.31	6.08	3.62	0.96	7.34	1	second
V21	30.10	32.63	11.56	0.21	0.33	0.07	2.72	0.71	4.33	6.28	3.47	0.43	7.15	1	second
V22	28.60	31.57	14.22	0.23	0.93	0.18	2.65	0.99	4.56	5.62	3.29	0.94	6.22	1	second
V23	35.28	27.31	11.89	0.21	0.14	0.18	2.84	1.07	4.12	5.33	3.66	0.72	7.24	1	second
V24	31.37	29.18	13.39	0.26	0.27	0.00	2.76	0.63	4.59	5.69	4.21	1.21	6.43	1	second
V6	35.14	30.01	10.87	0.83	1.14	0.41	3.28	0.60	3.38	2.55	4.09	0.76	6.94	2	first
V7	33.97	32.15	11.84	0.81	0.94	0.42	3.24	0.60	2.78	2.52	3.96	0.85	5.93	2	first
V8	36.31	30.52	7.77	0.99	0.79	0.41	3.59	0.76	2.84	4.57	4.36	0.56	6.53	2	first
V9	35.79	29.44	9.24	1.02	1.04	0.50	3.30	0.62	2.98	4.43	3.87	0.78	6.98	2	first
V10	33.85	28.79	11.93	1.11	0.93	0.50	3.24	0.78	3.25	4.48	4.31	0.66	6.15	2	first
V25	35.38	28.11	10.02	1.39	1.44	0.46	3.39	0.71	2.55	4.49	3.73	0.75	7.58	2	first
V26	33.00	28.43	11.19	1.41	1.37	0.47	3.64	0.49	2.97	4.61	4.23	0.74	7.45	2	second
V27	35.28	26.69	10.63	1.33	1.40	0.49	3.10	0.80	3.43	4.93	4.07	0.79	7.06	2	second
V28	34.57	28.75	11.56	1.30	1.25	0.44	2.96	0.39	3.04	4.63	3.53	0.48	7.10	2	second
V29	36.71	27.40	10.75	1.07	1.32	0.34	3.21	0.53	2.66	5.08	3.59	0.44	6.92	2	second
V30	33.16	28.80	11.20	1.33	1.23	0.44	3.07	0.52	3.50	5.76	3.91	0.91	6.16	2	second
V31	32.33	27.34	12.59	1.47	1.11	0.20	3.11	0.79	4.23	5.96	3.50	0.71	6.66	2	second
V32	32.96	28.77	11.34	1.35	1.11	0.46	3.27	0.78	3.28	5.22	3.54	0.64	7.27	2	second
V33	32.82	30.77	8.85	1.42	0.95	0.49	2.75	0.59	3.79	5.10	4.73	0.62	7.10	2	second
V34	29.29	29.82	12.41	1.27	0.94	0.75	2.63	0.65	3.18	6.02	4.92	1.02	7.11	2	second
V35	29.84	29.05	12.65	1.24	1.14	0.25	2.34	1.08	4.20	7.41	4.27	0.90	5.62	2	second
V36	31.76	27.60	12.12	1.28	0.47	0.69	2.81	1.14	4.68	5.13	4.21	1.44	6.69	2	second
V37	29.00	28.56	14.21	1.46	1.03	0.59	2.83	0.93	4.70	5.42	3.66	0.95	6.66	2	second
V38	28.90	30.20	13.87	1.52	1.27	0.32	2.83	0.74	3.98	5.06	3.98	0.82	6.52	2	second

Chapter 5. Conclusion

5.1 Dissertation Summary

The work reported in this dissertation mainly focuses on method development towards identification and quantification of glycopeptides or glycans derived from multiple glycoprotein samples via mass spectrometry-based (MS-based) strategies. MS-based strategies among other analytical methods, are frequently used in the field of glycomics analysis due to the higher sensitivity, higher resolution, and tandem capabilities provided by them.[1-3]

Protein glycosylation is one of the most prevalent post-translational modifications, which affects both the structures and the functions of the proteins.[4] However, this modification is sensitive to the changes of the cellular environment,[1, 5] resulting aberrantly glycosylated proteins during the progression of many diseases.[6-13] Therefore, altered glycans or glycopeptides derived from specific protein(s) of biological samples can serve as potential biomarkers for disease diagnosis and prognosis. However, the accurate identification and quantification of these heterogeneous glycans or glycopeptides is a challenge; it depend on not only the sensitivity of the analytical method, but also on the availability of effective sample preparation strategies and proper data classification approaches that handle multiple data inputs.

In glycomics analysis, development of effective strategies for glycosylation site profiling of complex glycoproteins is important; because, it allows not only the full characterization of the glycoproteins, but also the identification of site-specific glycan alterations that occur during the progression of diseases. The state-of-the-art expert analysis method for the glycosylation site identification typically requires higher initial glycoprotein amounts, two parallel LC-MS experiments, and higher data analysis time, which reduces the efficiency of the method.[14-15]

Therefore, chapter 2 introduces a rapid glycosylation site profiling method, which relies on high-resolution MS data and chromatographic retention times to track the glycosylation sites of proteins; the method identifies co-eluting glycoforms those extracted based on a pre-developed *N*-linked glycan library. The method successfully profiled glycosylation sites of a heavily glycosylated human plasma protein while identifying even the low abundant glycopeptides, those that could have gone undetected if the analysis was based on software tools. Therefore, the developed expert analysis approach, which relies on a single LC-MS run, while requiring half the analysis time and half the protein amount compared to the competing analysis method, would be useful not only to track the glycosylation sites of proteins, but also to identify low abundant glycopeptides that might represent an immunogenic glycoform of a biotherapeutic drug or a potential glycan biomarker for a certain disease.

Uromodulin, the most abundant protein excreted in human urine, is glycosylated and changes to its glycosylation profile have been reported during the progression of kidney diseases; thus, it can serve as a biomarker for kidney health.[9] Current methods for analyzing uromodulin glycosylation involve many time-consuming and laborious sample preparation steps, such as, glycoprotein enrichment, glycan labeling, and post-sample clean-up steps, all prior to the MS analysis.[9, 16-17] These complex sample preparation steps limit the current method’s applicability in readily assessing the uromodulin glycosylation changes in kidney-related studies. Therefore, resolving the challenges of tedious sample preparation steps involved in uromodulin glycosylation analysis is important; thus, chapter 3 introduces a clinically viable direct (-)ESI-MS method that allowed the quantification of *N*-linked glycans of uromodulin, extracted from human urine samples of two different biological states. The developed method omitted glycan labeling steps and post-sample clean up steps; it provided highly reproducible uromodulin glycan quantitation data over multiple analyses and over multiple sample preparations, while allowing glycomics samples of the same biological state to be classified together by PCA. Therefore, this method can be applied to quantify glycosylation differences of uromodulin samples derived from large cohorts of clinical samples of different biological states, followed by sample classification based on their glycosylation differences.

Sample classification based on a single glycan feature or a few selected glycan features is the typical approach; however, considering of all the individual glycan abundances of a sample along with their relationships to each other, to classify samples, can be useful in identifying the underlying trends that benefit the sample classification.[18] The Aristotle Classifier is one such newly developed data classification approach; it distinguishes glycomics samples based on their whole glycomic profile, instead of selecting one glycan feature or a few glycan features.[18] Once the classifier was built, it needed to be optimized; but, acquiring clinical glycomics samples with known glycosylation differences within them was a challenge. Therefore, chapter 4 describes an optimized sample preparation approach that we performed to chemically generate large sets of IgG glycopeptides data mimicking two different biological states, followed by the application of those data to the Aristotle Classifier and to a competing classification approach: PCA, to challenge and compare each approaches' classification abilities. The generated data proved to be optimum; they produced a challenging but tractable classification problem that allowed different algorithms' merits to be compared. The Aristotle Classifier outperformed the PCA classification, while showing its capability in successfully handling multiple data inputs. Therefore, the Aristotle Classifier can be used to distinguish glycomics samples from a large sample set; one good example is the classification of uromodulin samples derived from large groups of clinical samples. Furthermore, this classifier can be applied to other challenging classification problems where the competing PCA approach fails.

5.2 References

1. Dalpathado, D. S.; Desaire, H., *The Analyst* **2008,** *133* (6), 731-8.

2. Leymarie, N.; Zaia, J., *Anal Chem* **2012,** *84* (7), 3040-8.

3. Wuhrer, M., *Glycoconjugate J.* **2013,** *30* (1), 11-22.

4. Dwek, R. A., *Chemical reviews* **1996,** *96* (2), 683-720.

5. Zhu, Z.; Desaire, H., *Annual Review of Analytical Chemistry* **2015,** *8,* 463-483.

6. Saldova, R.; Reuben, J. M.; Abd Hamid, U. M.; Rudd, P. M.; Cristofanilli, M., *Annals of oncology : official journal of the European Society for Medical Oncology* **2011,** *22* (5), 1113-9.

7. Tsai, T. H.; Wang, M.; Di Poto, C.; Hu, Y.; Zhou, S.; Zhao, Y.; Varghese, R. S.; Luo, Y.; Tadesse, M. G.; Ziada, D. H.; Desai, C. S.; Shetty, K.; Mechref, Y.; Ressom, H. W., *Journal of proteome research* **2014,** *13* (11), 4859-68.

8. Zhang, X.; Wang, Y.; Qian, Y.; Wu, X.; Zhang, Z.; Liu, X.; Zhao, R.; Zhou, L.; Ruan, Y.; Xu, J.; Liu, H.; Ren, S.; Xu, C.; Gu, J., *PLoS One* **2014,** *9* (2), e87978.

9. Argade, S.; Chen, T.; Shaw, T.; Berecz, Z.; Shi, W.; Choudhury, B.; Parsons, C. L.; Sur, R. L., *Urolithiasis* **2015,** *43* (4), 303-12.

10. Russell, A. C.; Simurina, M.; Garcia, M. T.; Novokmet, M.; Wang, Y.; Rudan, I.; Campbell, H.; Lauc, G.; Thomas, M. G.; Wang, W., *Glycobiology* **2017,** *27* (5), 501-510.

11. Lundstrom, S. L.; Yang, H.; Lyutvinskiy, Y.; Rutishauser, D.; Herukka, S. K.; Soininen, H.; Zubarev, R. A., *Journal of Alzheimer's disease : JAD* **2014,** *38* (3), 567-79.

12. Ruhaak, L. R.; Uh, H. W.; Beekman, M.; Hokke, C. H.; Westendorp, R. G.; Houwing-Duistermaat, J.; Wuhrer, M.; Deelder, A. M.; Slagboom, P. E., *Journal of proteome research* **2011,** *10* (4), 1667-74.

13. Vanderschaeghe, D.; Meuris, L.; Raes, T.; Grootaert, H.; Van Hecke, A.; Verhelst, X.; Van de Velde, F.; Lapauw, B.; Van Vlierberghe, H.; Callewaert, N., *Molecular & cellular proteomics : MCP* **2018,** *17* (12), 2508-2517.

14. Wang, B.; Tsybovsky, Y.; Palczewski, K.; Chance, M. R., *Journal of the American Society for Mass Spectrometry* **2014,** *25* (5), 729-41.

15. Zhao, J.; Liu, Y. H.; Reichert, P.; Pflanz, S.; Pramanik, B., *Journal of mass spectrometry : JMS* **2010,** *45* (12), 1416-25.

16. Parsons, C. L.; Stein, P.; Zupkas, P.; Chenoweth, M.; Argade, S. P.; Proctor, J. G.; Datta, A.; Trotter, R. N., *The Journal of urology* **2007,** *178* (6), 2665-70.

17. Serafini-Cessi, F.; Bellabarba, G.; Malagolini, N.; Dall'Olio, F., *Journal of immunological methods* **1989,** *120* (2), 185-9.

18. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., *Anal Chem* **2019,** *91* (17), 11070-11077.