HIGH THROUGHPUT GLYCOMICS

by

MELODY PERLMAN PORTERFIELD

(Under the Direction of Mike Tiemeyer and Lance Wells)

ABSTRACT

Over half of all cellular proteins are modified by post translational addition of oligosaccharides. Proper glycosylation plays a critical role in cell-cell communication and changes in pH, nutrient availability, and cell status results in altered cellular glycosylation profiles which have been reported in a broad range of diseases including cancer, autoimmune diseases, and type II diabetes. Vaccination development relies heavily on differential recognition of glycan variability by the immune system and they are potential biomarkers for early detection of cancer. In addition, glycans play important roles in therapeutic applications, including both treatments and diagnostics.

Comprehensive characterization of the glycans on glycoproteins has become an essential element for drug development, quality control, and basic biomedical research. Manual interpretation of mass spectrometry datasets constitutes the core of most glycomics technology currently in use. However, interpretation of up to 2000 mass spectra per biological sample consumes significant expert personnel time and reduces the number of samples that can be analyzed. This bottleneck is a major impediment blocking the expansion of glycomic analysis to a broad range of basic biomedical investigations. Progress in the field has been severely restricted by the absence of appropriate computational software tools that facilitate automated structural assignment and high throughput data analysis.

We have combined efforts of computer scientists, experimentalists, and mass spectrometrists in an effort to provide a semi-automated high throughput workflow aimed to fulfill this critical need.

INDEX WORDS: glycan, mass spectrometry, glycomics, automation, software, throughput, pancreatic cancer, biomarker

HIGH THROUGHPUT GLYCOMICS

by

MELODY PERLMAN PORTERFIELD

B.S., Emory University, 1995

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

HIGH THROUGHPUT GLYCOMICS

by

MELODY PERLMAN PORTERFIELD

Major Professors:	Lance Wells
Committee:	Mike Tiemeyer I. Jonathan Amster

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

DEDICATION

To my parents, Jim and Linda Murphy: I am forever grateful for your endless support. You have always believed in me, and given me more than I deserved, I am blessed and grateful to be your daughter.

To my husband, Michael Porterfield: Thank you for believing in me, especially when I didn't. Your patience, love, and "no job is too big or hard" attitude have been a continual source of strength for me. You are and will always be my happily ever after.

To my children: Although I have relentlessly strived and prayed that I was able to teach you all the important things in life, it has turned out that it was you who have taught me the most. Hunter, you are strong, smart, loyal, and my first true love. Keep your heart open and you will reach the stars. Gavin, you are kind, loving, and the blessed glue that holds us all together. Know that you can do anything you dare to dream. Ava, you were just an infant when I began this work; you have grown into a beautiful girl. You have a passion and endless curiosity that can move mountains. Never be afraid and shoot for the stars.

I love you all.

ACKNOWLEDGEMENTS

I would like to thank Mike Tiemeyer for believing in me and giving me this amazing opportunity, along with the tools and guidance to succeed in anything. You have truly led by example, passing on your seemingly endless body of scientific knowledge and more importantly your professionalism. I am forever grateful. To Lance Wells and Jon Amster, thank you for always reminding me to focus on the analytical chemistry and bringing out my best work. To Kazuhiro Aoki, my much older brother, dear friend and cubby mate: you have taught me everything I know about glycan analysis and I am forever grateful to you. To Rene Ranzinger, Will York, Brent Weatherly, and Khalifeh Al Jadda, without you this project would be nothing more than a pipe dream, thank you for making it a reality. Your collective efforts, expertise, patience and trust in me have enabled a tremendous amount of growth, for the field of glycomics and me personally. Thank you.

I would like to thank the University of Georgia Chemistry department and the Complex Carbohydrate Center for all the opportunities, professional development and knowledge you have provided. I would also like to thank all of my lab members for your helpful advice, patience, and family atmosphere that made this lab a real home away from home.

Last but not least, I would like to thank my family and friends who have been there along the way. I could not have done it without you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Along with nucleic acids, proteins, and lipids, complex polymeric sugar molecules known as glycans are one of the four major classes of macromolecules fundamental to all living systems. [1]. Glycosylation is the most abundant and structurally diverse class of protein post-translational modifications (PTM) and the complete set of glycan modifications present in a cell, tissue, or organism is referred to as the "glycome." The information content afforded by the addition of glycans to proteins exceeds that of any other class of molecules and substantially increases the diversity of phenotypes that are possible from a limited genotype. Compared to the study of proteins and nucleic acids, relatively little attention has been paid to glycans despite the critical roles they play in most biochemical processes which are fundamental to life on Earth.

All cells are coated with a dense layer of glycans, and all cellular interactions take place in the context of this layer including microbial attachment and entry, cell-cell adhesion, ligand receptor binding, and metastasis. The glycome is not only structurally more complex and diverse than the genome, proteome and transcriptome, it is also more dynamic, changing rapidly in response to environmental factors including metabolic, disease and developmental states. Glycans play roles in a wide variety of cellular processes including cell signaling, immunity, inflammation, and molecular recognition. Furthermore, alteration of glycans has been implicated in the etiology of every major disease known to man. Expanding our knowledge of glycans' structures and how they function in human health and disease can serve as a guide for intelligent design strategies for the detection, identification and treatment of diseases. Two promising areas include glycan based biomarker discovery for early cancer/disease detection and monitoring of treatment efficacy and development of new strategies to fight infection through host-pathogen interaction interventions as well as potential immune response modulation for infection and auto-immune/allergic conditions. Additionally, Development of glycan based small molecule and glycan modified biologic therapeutics is a thriving and promising multi-billion dollar industry. Full and comprehensive characterization of the glycans on glycoproteins has become an essential element for drug development, quality characterization, and basic biomedical research.

Chapter 3 investigates potential biomarker identification through a combined proteomic and glycomic analysis of human pancreatic ductal fluid. A comprehensive analysis of the N-linked glycome of pancreatic ductal fluid identified unexpected clustering of patient samples into discrete subgroups that are enriched in sialylation or fucosylation, or are mixed with respect to both types of glycans independent of diagnosis. Within each group, changes in glycan prevalences are detected comparing normal to cancer albeit on a small number of samples. But, across groups, the glycan expression changes are different, even opposite in some cases. Peng Zhao in Lance Wells's laboratory observed changes in expression levels and heterogeneity of secreted pancreatic enzymes and non-enzyme pancreatic proteins in cancerous ductal fluid in comparison to normal ductal fluid in a small training set of samples by LC/MS/MS. Therefore, interpretation of glycomic and glycoproteomic profiles must consider the heterogeneity of glycosylation across human populations before assessing the meaningfulness of changes in candidate biomarkers. The proteomic and glycomic features extracted from the training set of samples reported here establish important parameters for expanded validation and emphasize the need for large sample sets.

Mass spectrometry has become a tool of choice for analysis of biomolecules given the limited sample often available. Ease of peptide identification using tandem MS is mainly due to the linear structure of peptides and the availability of reliable peptide sequence databases. In proteomics, fragments with high intensities come from complete ion series, with the difference between two adjacent peaks representing the mass of an amino acid, thereby eliciting the amino acid sequence using the ion series. In glycomics MS data there are almost no complete ion series and the branched molecules further complicate sequence determination. In addition, isomeric monosaccharides share the same mass and differing only by the position of a hydroxyl group makes it impossible to distinguish them by MS alone. Furthermore, databases for glycans exist but are limited, minimally curated, and suffer pollution from glycan structures that are unlikely to be produced in nature or are irrelevant to the organism of study.

Advances in mass spectrometry instrumentation over the past decade have resulted in increased sensitivity, speed and mass accuracy of glycans released from biological samples. Consequently, datasets have also increased and regularly produce over 2000 mass spectra per sample when an MS/MS approach is taken and could produce tens of thousands of spectra that must be interpreted depending on the level of fragmentation desired.   Glycomic datasets rely on primarily on manual interpretation and requires significant personnel time and expertise which in turn reduces the number of samples that can be analyzed. This bottleneck is a major impediment blocking the expansion of glycomic analysis to a broad range of basic biomedical investigations. Progress in the field has been severely restricted by the absence of appropriate computational software tools that facilitate automated structural assignment and high throughput data analysis. A major aim of this work is to provide tools for the community that simply data interpretation and thereby enable glycomic analysis for a wide variety of biomedical investigators.

Glycomic datasets presented in Chapter 3 were manually interpreted over a period of 24 months and put a spotlight on the need for tools capable of automating and speeding up this task. Chapter 4 was born directly out of needs identified while undertaking the task of manually describing the glycome of a multitude of sample sets. The workflows and data analysis tools described are conservatively expected to shift throughput from one sample every three months to six samples every week. This acceleration of data analysis constitutes a paradigm shift in the field of glycomics by making statistical confidence available to investigators through increased sample number. The tools and workflows are a starting point and a work in progress that will facilitate rapid growth in the field of glycomics. Our goal is to make glycomic analysis a routine, albeit technically demanding, option for a broad range of biomedical investigations.

CHAPTER 2

LITERATURE REVIEW

Introduction

All cells are coated with a complex layer of covalently attached sugar chains known as glycans. Over half of all cellular proteins are glycosylated [2]. Proper glycosylation plays a critical role in development and differentiation of cells, cell-cell and cell-matrix interactions, host-pathogen interactions, fertilization, and signaling pathways. Cellular glycosylation profiles are altered under changing conditions including pH, nutrient availability, and cell status. Glycans are often large and can dominate the physiochemical properties of their carrier, affecting solubility, half-life, immunogenicity and biological activity[3].

Decades of research have shown that glycans are not only involved in normal physiology but also in the etiology of all major human diseases, both chronic and acute. Glycosylation changes have been reported in a broad range of diseases including cancer, autoimmune diseases, type II diabetes as well as the CDG's (congenital disorders of glycosylation) [4]. Glycans play a clear role in regulation of the immune system which responds to perceived danger as either inflammation or immunity which are both regulated by glycans as well[5]. Glycans regulate inflammation directly in a multitude of ways and inflammation underlies diabetes, arthritis, heart disease, asthma, and cancer. Antibodies which are themselves glycosylated are produced to fight pathogens which were detected by the presence of foreign glycans present on pathogens during infection.

In fact glycans dominate the interface of all self versus non-self recognition events including, pathogens, toxins, and erroneously identified targets of autoimmune diseases[6].

Influenza was responsible for 4 major human pandemics since the 1900's killing more than 50 million people and concern for new variants is a constant threat. Influenza infection begins when hemagglutinin (HA), a viral coat protein binds a glycan structure, sialic acid, on a host cell. Neuraminidase (N) cleaves the sialic acid which releases newly replicated virus particles from the host cell which then go on to infect other cells[7]. Tamiflu™ and Relenza™ are antiviral medications that target the neuraminidase and block influenza replication[8]. Different variations of HA bind different glycan epitopes preferentially and is the basis of species selectivity as in Avian (2-3linked sialic acid) vs human (2-6 linked sialic acids) influenza[9].

The HIV virus is another example illustrating the importance of glycans in human health and disease. The HIV gp120 viral coat protein responsible for binding the CD4 receptor on T cells to gain entry is coated in glycans[10]. HIV uses the glycans to disguise itself and evade attack by the host immune cells. In addition, HIV rapidly changes the glycans presented to prevent the immune system from forming effective antibodies capable of binding and neutralizing the virus and contribute to the poor performance of candidate HIV vaccines to date[11].

Autoimmune diseases can occur when antibodies are made in response to a pathogen that bears glycans that are similar to the host organism's glycans. For example, infection by *Campylobacter jejuni* can result in Guilian Barre syndrome, a life threatening disorder that affects the peripheral nervous[12]. Tissue damage occurs when the host immune system produces antibodies which mistakenly targets the host's nerve tissue which bears glycan similar to those presented by the bacterium. This type of auto-antibody event is also suspected in rheumatoid arthritis and systemic lupus as well.

Glycosylation changes have been reported in all types of cancer cells and many tumor associated antigens are glycans usually only seen on developing embryos. Many of the existing tests for cancer rely on differential glycosylated proteins. Cancer antigen 125 (CA125) and prostate-specific antigen (PSA) are two examples of glycoproteins used for detection and monitoring treatment efficacy.   Research has shown that core fucosylated alpha-fetoprotein (AFP-L3) improves diagnosis of hepatocellular carcinoma over standard AFP highlighting the importance of glycomic studies.

Glycans play important roles in therapeutic applications, including both treatments and diagnostics, some of which include: increased accuracy in diagnostics (example AFP), influenza transmission inhibition by neuraminidase treatment (Relenza™[8b] and Tamiflu™[8a]), therapy for osteoarthritis by hyaluronic acid, and a 3 fold increase in the half-life of erythropoietin (EPO) by addition of two glycosylation sites. Heparin, a polysulfated GAG that acts as a blood anticoagulant, is one of the most useful drugs in medicine today[13]. Vaccination development relies heavily on differential recognition of glycan variability by the immune system including type b influenza, *leishmania*, HIV, *Neisseria meningitides*, and *meningococcus*[14]. Glycans are also potential biomarkers for early detection of cancer[4, 15].

Considering that glycans significantly influence the functions of cellular proteins and the therapeutic potency of glycoprotein biologicals, ongoing efforts to develop robust, sensitive, and quantitative glycan characterization methods are well justified. Full and comprehensive characterization of the glycans on glycoproteins has become an essential element for drug development, quality characterization, and basic biomedical research. Manual interpretation of mass spectrometry datasets constitutes the core of most glycomics technology currently in use. However, Interpretation of up to 2000 mass spectra per biological sample consumes significant expert personnel time and reduces the number of samples that can be analyzed. This bottleneck is a major impediment blocking the expansion of glycomic analysis to a broad range of basic biomedical investigations. Progress in the field has been severely restricted by the absence of appropriate computational software tools that facilitate automated structural assignment and high throughput data analysis.

Classification of Glycans

Major glycan classes are defined by the carriers they are linked to, protein or lipid, as well as the type of linkage, N-linked or O-linked (Figure 2.1). N-glycans are covalently linked to proteins with the consensus sequence Asn-X-Ser/Thr at an asparagine residue by formation of an N-glycosidic bond with a GlcNAc residue. N-glycans usually have a common core structure of $Man_3GlcNAc_2$ called the trimannosyl core. Subsequent elongation of the trimannosyl core generates three classes of glycans, high mannose, complex, and hybrid. The most common O-linked glycans, mucins, are covalently linked to the –OH of a serine or threonine by formation of a glycosidic bond with a GalNAc residue. O-glycans are generally smaller and less branched than N-glycans, and exhibit several different core structures including cores 1-8 for mucins, O-Man, O-GlcNAc, O-Fuc, and O-Glc; much diversity exists in this class. Glycosyltransferases extend N- and O-glycan cores by the stepwise addition of monosaccharide subunits. Common occurring additions include neolactosamine units

(Galβ1-3GlcNAc) and lactosamine units (Galβ1-4GlcNAc), both of which are often elaborated further by sialic acids and fucose, as well as blood group and Lewis family specific sugar epitopes. Proteoglycans have one or more linear sulfated glycan (GAG) chains linked to a core protein at a serine or threonine residue by formation of a glycosidic bond with a xylose subunit. Other classes of glycoconjugates include GPI anchors, glycosphingolipids, and glycoglycerolipids. The Consortium for Functional Glycomics (CFG)[16] has standardized abbreviations and cartoon representations for glycans for consistency and ease of illustration of glycan moieties[15].

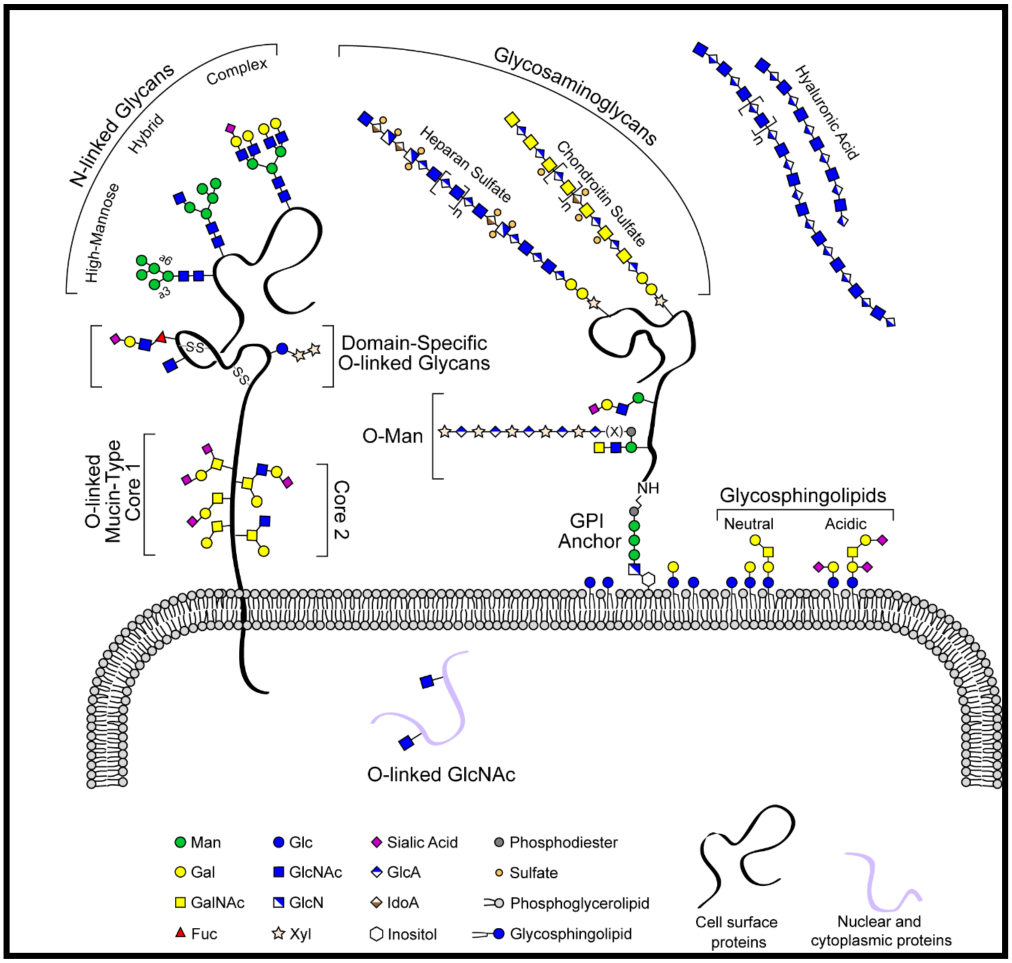


Figure 2.1 Classes of Glycans
(adapted from Moremen, Tiemeyer, and Nairin 2012)[17]

All monosaccharides are made up of a chain of chiral hydroxymethylene units with a hydroxymethyl group at one end and an aldehyde or ketone group on the other end. The aldehyde carbon is labeled C-1 whereas the carbonyl group in ketoses is at C-2. The absolute configuration, D or L, is determined by orientation of substituents at the highest numbered asymmetric carbon. When in its cyclic state, a hemiacetal group is formed by the reaction of one of the hydroxyl groups with C-1. Both 5 (furanose) and 6 (pyranose) member rings can be formed, C-(1or2)-O-C-4 and C-(1or2)-O-C-5 respectively. A glycosidic bond is formed between the anomeric carbon (usually C-1) of one monosaccharide and a hydroxyl group of another, (i.e. a hemiacetal group reacts with an alcohol to form a full acetal). Linkage anomerocity, α or β, is determined by the side of attack, where α indicates the bond is below the plane of the monosaccharide and β indicates it is above it. The end of the sugar chain that is (or was) attached to the peptide is known as the reducing end whereas the end furthest away is referred to as the non-reducing end[15]. Glycan structure fragments are named according to Domon and Costello[18] nomenclature (Figure 2.2).

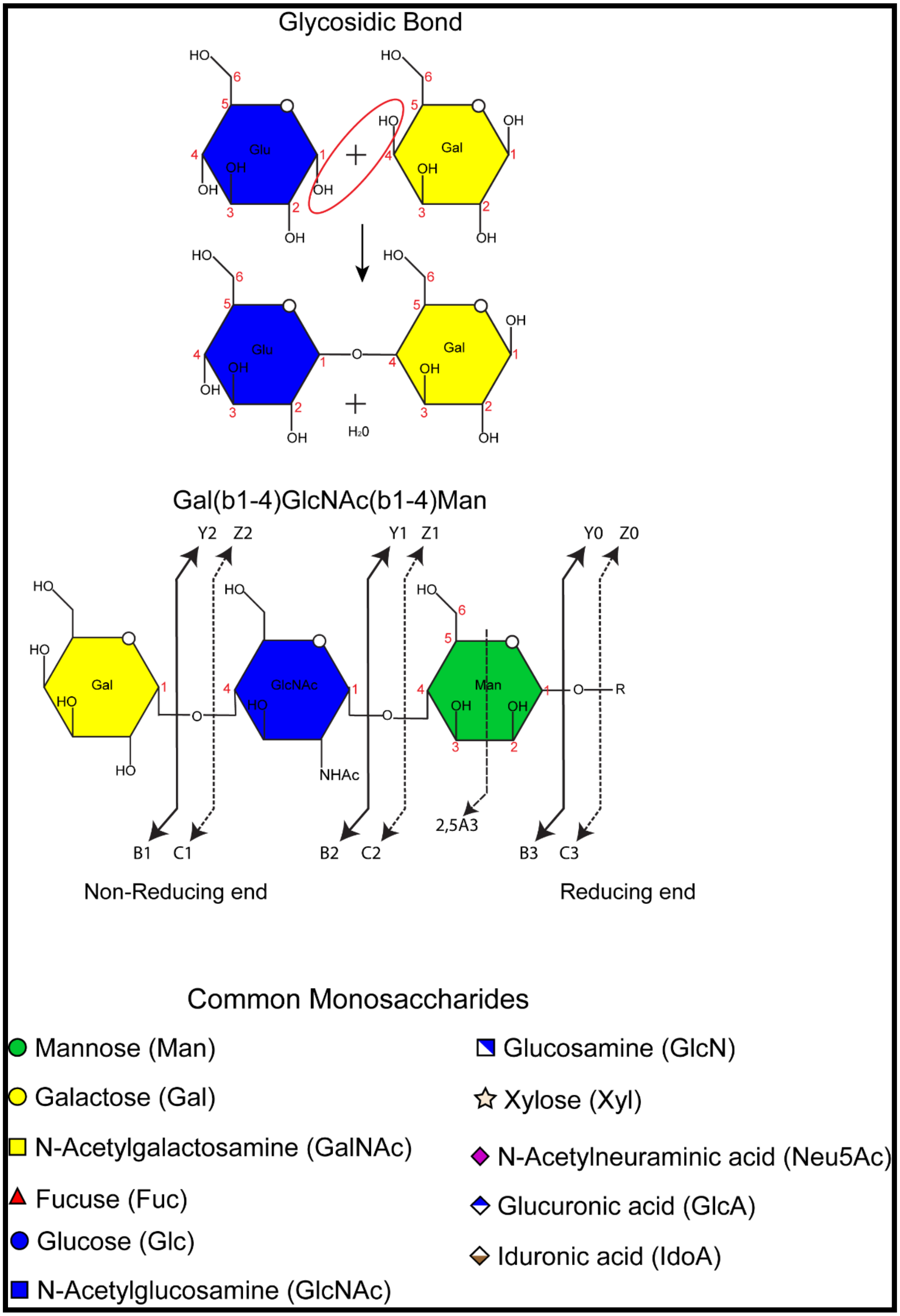


Figure 2.2 Monosaccharides, glycosidic bond and fragmentation nomenclature

Although hundreds of distinct monosaccharides occur in nature, only ten of them serve as the major building blocks for all human glycopeptide glycans. They include D-glucose (Glc), D-mannose (Man), D-galactose (Gal), L-fucose (Fuc), D-N-acetylglucosamine (GlcNAc), D-xylose (Xyl), D-glucuronic acid (GlcA), D-N-acetylgalactosamine (GalNAc), and the sialic acids primarily N-acetyl-neuraminic acid (Neu5Ac) and N-glycolyl-neuraminic acid (Neu5Gc)[19]. Except for the sialic and uronic acids, the hexose or hexosamine building blocks are very similar in molecular weight and charge, making them difficult to distinguish. Furthermore, the complexity of topologies present in glycans makes the task of unraveling their structure especially difficult. However, this complexity provides the diversity important for glycans to govern many aspects of the cell's processes.

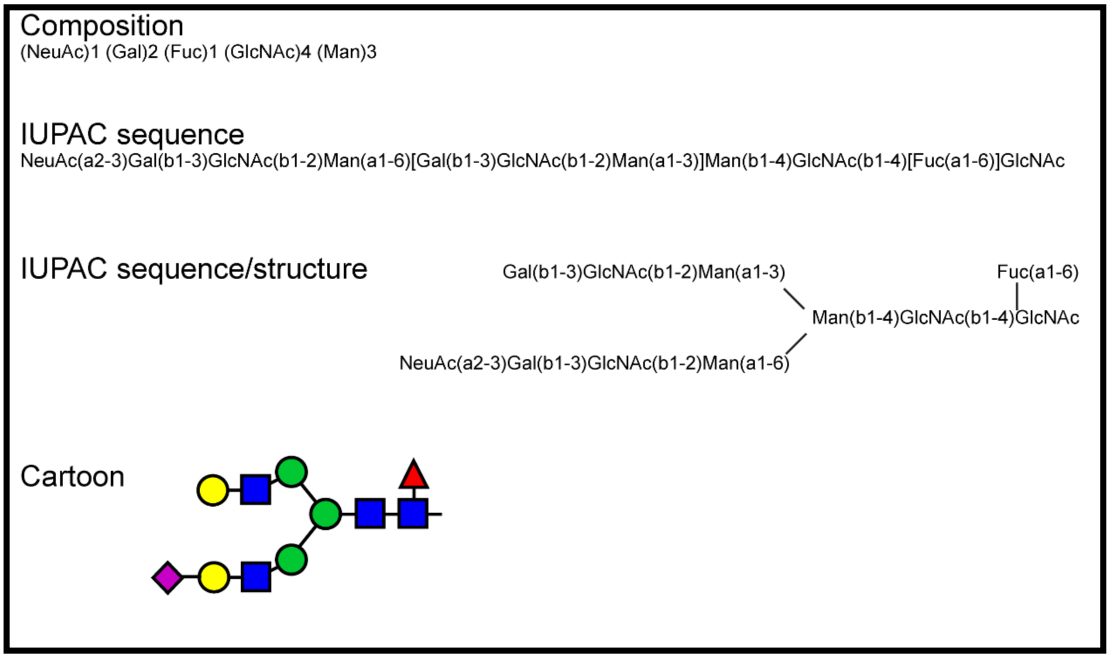

Composition
(NeuAc)1 (Gal)2 (Fuc)1 (GlcNAc)4 (Man)3
IUPAC sequence
NeuAc(a2-3)Gal(b1-3)GlcNAc(b1-2)Man(a1-6)[Gal(b1-3)GlcNAc(b1-2)Man(a1-3)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc
IUPAC sequence/structure
Gal(b1-3)GlcNAc(b1-2)Man(a1-3)
Fuc(a1-6)
Man(b1-4)GlcNAc(b1-4)GlcNAc
NeuAc(a2-3)Gal(b1-3)GlcNAc(b1-2)Man(a1-6)
Cartoon

Figure 2.3 Glycan structural representations

Glycan synthesis occurs in the endoplasmic reticulum and Golgi apparatus where a combination of the roughly 250 specific enzymes involved in glycan synthesis act in a highly regulated way[20]. Differential glycan expression is controlled by varying combinations of glycosyltransferase/exoglycosidase expression and activity which is dictated by the cell's type, status, and environment, although exact mechanisms are not completely understood[21]. Unlike the linear nature of DNA, RNA, and protein synthesis, monosaccharide subunits are added sequentially with branching points and anomeric configurations forming a tree-like topology. Three amino acids or nucleotides can be combined in six possible sequences whereas three hexoses alone can be theoretically combined to form 1,056 different sequences[22]. Furthermore, the lack of a genetic template from which to predict glycan sequences renders database searches comparing experimentally derived fragmentation data with a reference library for rapid data analysis as is done in proteomics unlikely.

Analytical Strategies

Due to the chemical similarity of monosaccharide subunits and the complexity of sequence and linkage possibilities that glycans possess, no single analytical method is capable of complete compositional and structural determination of biological samples with limited quantities. Detailed analysis requires a combination of approaches.

Glycans can be released from their carrier enzymatically or chemically depending on the class of glycans present as well as the type of analysis desired. The process of acid hydrolysis can be used to cleave all glycosidic bonds reducing glycans to their monosaccharide building blocks to give compositional analysis[23]. Hydrazinolysis or β-elimination using alkali treatment can be used to cleave intact glycans from their carriers and can be optimized to target N- or O-glycans or a mixture of both[24]. Intact glycans can also be cleaved from the peptide backbone by endoglycosidase enzymes including PNGase F or PNGase A for N-glycans[25] as well as Endo H for high mannose and hybrid N-glycans specifically[26]. Sequential exoglycosidase digestion of glycans produces data capable of differentiating glycan composition, sequence and linkage anomerocity[26b]; however, large quantities of sample and time are required.

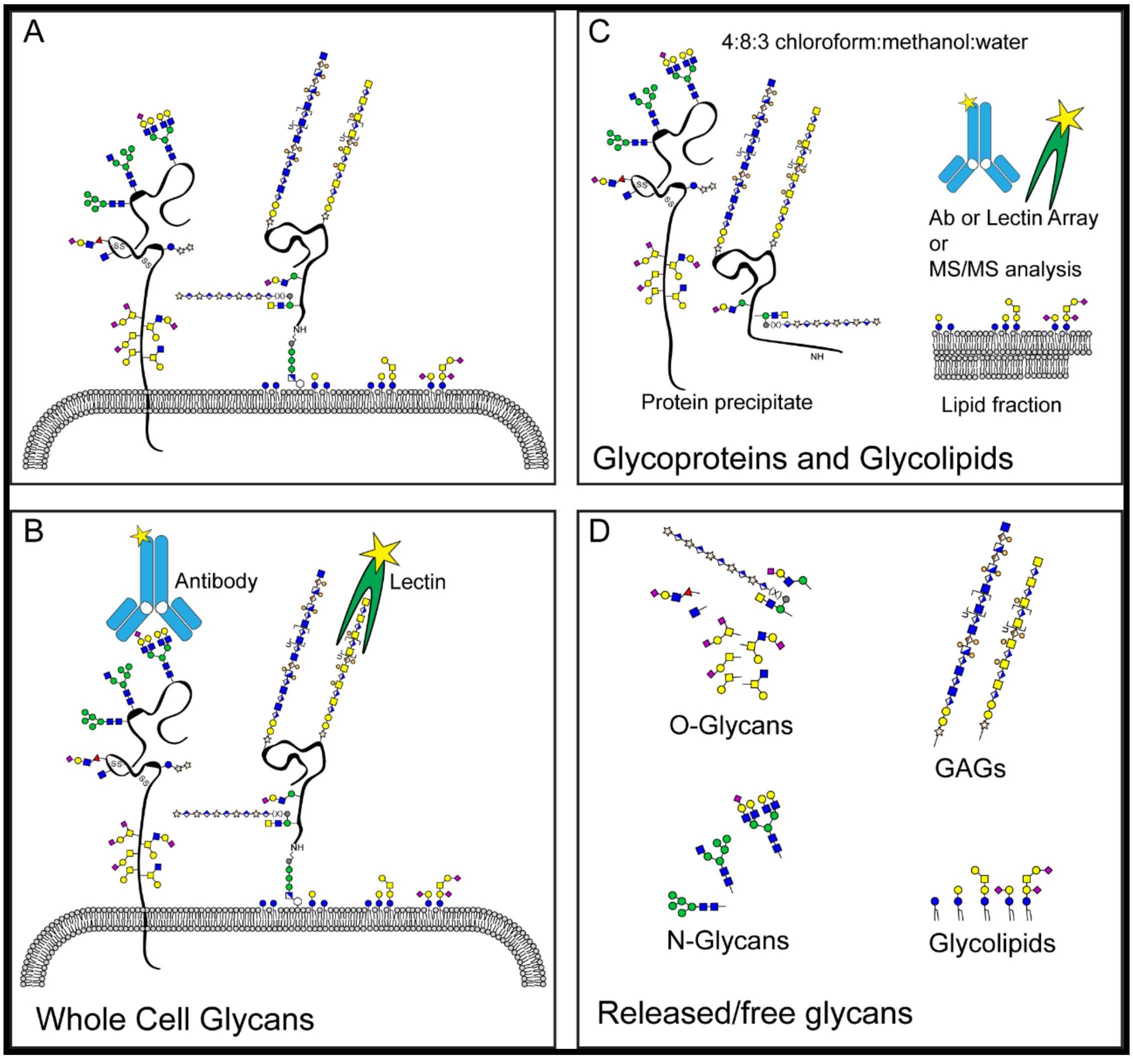


Figure 2.4 Multiple approaches for profiling glycans

Released glycans have a very low extinction coefficient in the UV range, limiting the sensitivity of direct detection by UV-vis methods. However 10-100 pmols can be seen by pulsed amperometric (PAD) or refractive index detection[27]. Sensitivity of detection to the sub-picomolar range can be achieved by addition of a fluorescent tag such as 2-AB, 2-AP, 2-AA or ANTS by reductive amination[28]. It is important to note that all labeled glycans produce the same molar response, which is crucial for absolute quantification.

Enzymatic or chemically released glycans can be chromatographically separated by a wide range of techniques, including normal [26a]and reverse phase HPLC separations based on hydrophobicity differences, as well as capillary electrophoresis (CE), capillary affinity electrophoresis (CAE), and weak anion exchange (WAX) HPLC separations based on charge differences[29]. In addition, fluorophore-assisted carbohydrate electrophoresis (FACE or PAGEFS) [30]separates fluorescently labeled glycans electrophoretically on polyacrylamide gels using equipment widely available in most labs. Gas chromatography can also be used to separate monosaccharides and small glycans after appropriate derivitazation (PMAA, TMS)[29].

Methods of detection and identification of pooled or separated glycans include the use of lectins for visualization, retention time measurement in chromatography, molecular weight determination in mass spectrometry, and measurement of magnetic field distortion in nuclear magnetic resonance (NMR) techniques. Each of these methods can provide different pieces of the puzzle which can create a comprehensive picture of a glycan’s structure when reassembled correctly.

Over 150 different lectins are commercially available[3] and have been used to isolate and visualize glycans in a variety of ways including conjugation with biotin, HRP, and fluorescent tags[31]. While some lectins recognize different glycans quite selectively, very few of them have been fully-characterized. Lectins offer a speedy, cost effective, high throughput method of glycan analysis and have been used to map glycosylation patterns and when immobilized on a stationary support facilitate affinity purification and enrichment. However, accuracy of the representation of glycan abundance by lectin arrays in native physiological states has yet to be demonstrated.

Detection of glycans by fluorescence measurement, as is often used in HPLC methodologies, is extremely sensitive. Glycans can be separated chromatographically and fractions collected for further analysis as well. This approach is the only methodology currently available for absolute quantification of a broad range of glycans[32]. Identification of glycan structures is based on comparison with retention times of known standards, which can be useful although limited by the inability to identify novel structures inherent to all methods dependent on standard libraries. Structural resolution and throughput are limited with HPLC methods but can be increased with expertise[32-33]. Quantitative compositional analysis of mono and oligosaccharides from simple mixtures and purified samples is commonly carried out by this method[5].

NMR is the only technique that can singlehandedly yield a full structural picture of a glycan, including linkage anomerocity, sequence and branching details as well as full monosaccharide identification[24]. The drawbacks of this technology are that the amount of sample required is not obtainable from most biological materials, and that expensive equipment and technical expertise is required[29]. Much can be learned about glycan-glycan and glycan-protein interactions using NMR. Because it provides structural information with reasonable sensitivity compared to NMR, mass spectrometry has become a tool of choice for the detection and identification of glycans from biological samples.

Table 2.1 Summary of glycan analytical approaches

	NMR	**HPLC**	**MS**	**MS/MS**	**MSn**
Glycan label	Native	Fluorescence derivatization, 2AB, 2AP, or native	Permethylation (most common), 2AP, 2AB, or native	Permethylation (most common), 2AP, 2AB, or native	Permethylation (most common), 2AP, 2AB, or native
Advantages	Complete structural details, composition, conformation & linkage	Absolute quantification, sensitivity, separation/fraction collection possible, inexpensive	Speed, coverage, ease of sample handling & data interpretation, software availability, sensitivity	Speed, coverage, ease of sample handling, some structural data obtained	Structural data rich-differentiate isobaric structures, some linkage and anomerocity details obtained
Dis-advantages	Large sample size required and high level of expertise needed, high cost	Relies on standards, not good for complex mixtures, structural details limited, co-eluting compounds	No structural data obtained, identification on intact m/z, isobaric structures not differentiated, dimer/trimer formation	Structural data limited, spectra interpretation manual	Data interpretation difficult and time consuming
Structural Resolution	Very high	Low	Low	Medium	High
Throughput	Low	Low/Medium	High	Medium	Low
Expertise level required	Very high	Medium	Low	Low/Medium	High
Application	Glycan-glycan or glycan-protein interactions	Quantitative compositional analysis of mono & oligo saccharides from mixtures	Rapid general glycan profile comparison between different samples (cancer vs. noncancer)	General glycan profile comparison between different samples with some structural data support	Specific epitope identification/ differentiation ex. Lex for biomarker discovery, isobaric structure differentiation

Mass Spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass of all molecule's present in a sample[34]. Measured masses are reported as a spectra where mass/charge (m/z) is plotted on the x-axis versus relative abundance on the y-axis. Intact molecules measured in a full MS or can be isolated and broken into smaller pieces (MS/MS) and the resulting fragments can be measured in a similar fashion. Initially MS was primarily used for small molecules, however, advances in sensitivity, robustness, and resolution have paved the way for a rapid expansion in development of applications for biological molecules, specifically, identification and quantification of peptides and most recently glycans[35]. In order to measure molecules, they must first be ionized, sorted and separated according to their mass and charge. All mass spectrometers consist of three basics components, an ionization source, a mass analyzer, and an ion detector, however multiple types exist for each component of the instrument and can be combined in varying ways to obtain the desired results[34].

While many types of ionization exist, the two techniques best suited for large biomolecules are the soft ionization techniques known as Matrix-Assisted Laser Desorption Ionization (MALDI) and Electrospray Ionization (ESI). In MALDI, analytes are first co-crystalized in an organic matrix solution such as 2,5-dihydroxy benzoic acid (DHB) which serves to protect the analyte from destruction[36]. Analyte/matrix mixtures are vaporized when pulses of UV laser beam at a fixed wavelength rapidly heat the mixture. Ionization of the analyte occurs when charge is transferred from the matrix to the analyte causing the analyte to gain or lose a proton (positive mode vs negative mode).

In ESI, analytes are dissolved in an appropriate solvent containing a selected salt, and passed through a heated capillary[37]. A high voltage electric field is applied as solvent/analyte droplets are sprayed causing them to break apart into a fine mist. Heat from the capillary combined with a flow of Nitrogen gas causes the solvent to evaporate leaving the analyte carrying one or more charges[38].

MALDI and ESI each have advantages and disadvantages that must be considered. The main advantage of MALDI is that most molecules carry only one charge therefore producing a straight forward spectrum[39] when compared to that of multiply charged ESI spectrum where analytes are represented at multiple mz's which can be difficult to read. MALDI is suited for very large molecules in excess of 300,000 atomic mass units (amu) where ESI is limited to 100,000 amu's however MALDI suffers in the lower mass range below 600 amu due to interference from matrices[40]. Unlike ESI, MALDI is tolerant to salts, buffers, and detergents making sample handling and preparation easier. The main disadvantage of MALDI, especially when analyzing glycans, comes from the presence of labile bonds that can be easily broken during this type of ionization, for example sialic acid modifications are often disrupted.

In addition to different types of ionization strategies, many different types of mass analyzers are available each with their own set of advantages and limitations as well. Mass analyzers all perform the task of essentially weighing analytes of interest but have varying approaches. All mass analyzers operate under a high vacuum to keep ions from colliding and can be paired with different ionization techniques and methods of detection. Key aspects to consider for comparing mass analyzers include their mass accuracy, resolution, and sensitivity. Mass analyzers can be grouped into time-of-flight, quadripole, ion traps, and Fourier transformation including both magnetic sector and orbitrap methodologies.

Time-of-flight (TOF) mass analyzers operate on the principle that upon simultaneous ionization and introduction into the flight tube, small ions travel faster and reach the detector before larger ones[39]. TOF instruments measure the time it takes for ions to travel and converts that to mass. TOF is most often paired with MALDI since all ions must be introduced simultaneously. TOF is fast, sensitive, and suitable for a wide mass range, however it suffers from relatively low resolution when compared with others[25].

Quadrupole mass analyzers use a combination of RF and DC voltages applied to four parallel metal rods to act as a mass filter. Different frequencies allow masses to pass through the length of the quadripole one at a time before exiting to the detector. The quadripole can select one mass at a time or scan through the entire mass range using scanning mode[41]. Multiple quadrupoles can be connected in sequence such as in the triple quad (QQQ) instrument to achieve high selectivity and sensitivity especially for low abundance analytes, however this type of mass analyzer configuration suffers from low resolution. Additionally, when a complete spectrum is required as in scanning mode, sensitivity is greatly reduced as well[42].

Ion trap (IT) mass analyzers are similar to quadrupole analyzers, however, a ring electrode rather than metal rods form a 3 dimensional space where ions are trapped and stored. RF and DC voltages are applied focusing ions in a small volume where they can be selectively excited and become unstable causing them to be ejected to the detector. The linear ion trap (LTQ) is a variation that traps ions in a 2 dimensional space that has superior trapping capacity, resolution, and rapid ion accumulation. Ion traps have the advantage of not only being capable of measuring intact analytes, but can also break them into fragments and re-measure the pieces. Ion traps are limited by capacity of the trap, which means there is a maximum number of ions that will fit in the trap without effecting the behavior of each other and producing spectral distortions due to space charging effects and sensitivity/dynamic range suffers.

Two types of mass analyzers rely on measuring the frequency in which analytes travel in orbit in response to a magnetic field as in Fourier Transform Ion Cyclotron Resonance (FTICR) or an electrostatic field as in the orbitrap mass analyzer. Both analyzers trap the ions and allow them to oscillate and repeatedly measure their frequencies which are a function of their mass-to-charge ratio. Fourier transformation deciphers the complex and overlapping transient trace of analyte motion into mass spectra. These instruments have the highest mass accuracy and resolution available but are limited in sensitivity due to space charge limitations as well[43].

All samples in this thesis were analyzed using an LTQ-Orbitrap Discovery hybrid mass spectrometer[44] that incorporates an LTQ XL linear trap and the Orbitrap manufactured by Thermo-Fisher. It was equipped with nano-electrospray ionization (NESI). Permethylated glycans were dissolved in 1mM sodium hydroxide in 50% methanol, and infused into the Orbitrap. High resolution full MS were taken in FT mode in the Orbitrap and fragmentation by collision induced dissociation (CID) occurred in in the ion-trap. Manufacturer specifications are reported as maximums[45]: Resolution 30,000 FWHM at 400 m/z, mass range of 50-2000 m/z, mass accuracy 5ppm with external calibration, dynamic range 4,000 in a single spectrum and 10,000 between spectra[46].

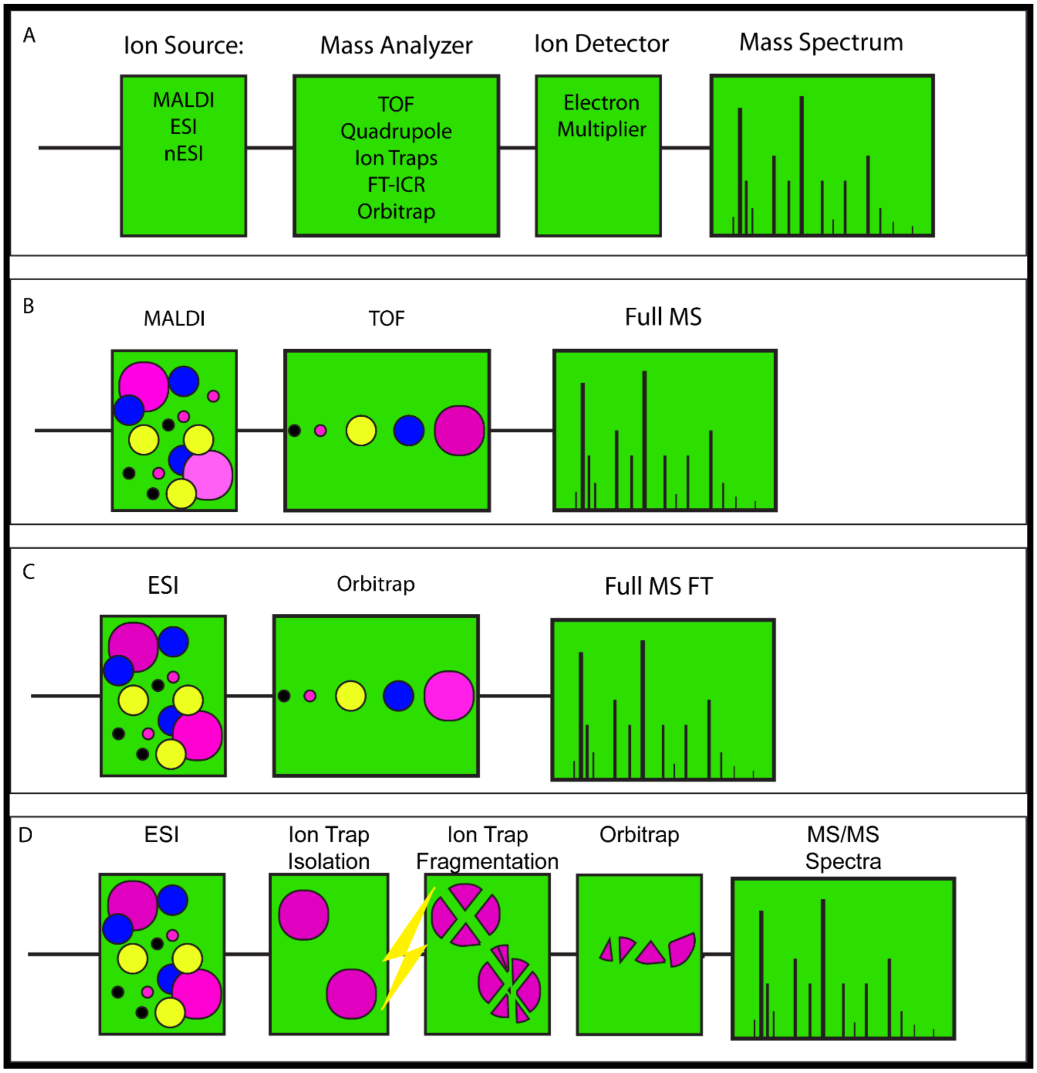


Figure 2.5 Mass Spectrometry variations

Mass Spectrometry Data Acquisition

There are two basic strategies in MS based glycan analysis, single-MS, in which a full MS spectrum is generated yielding an accurate mass to identify and relatively quantitate intact glycans in a sample, and MS^n, which adds one or more dimensions of mass analysis by fragmentation in order to obtain structural details. A continuum exists in the scientific community regarding the acceptable level of detail required to characterize glycans. On one end, total molecular weight may be sufficient for some investigators and at the other end, identification of each linkage position and type may be required. Each approach has its advantages and disadvantages depending on the goal and needs of the investigator.

The single-MS approach is usually paired with MALDI-TOF technology and has the advantages of speed, broad range of mass coverage, and straight forward data interpretation. Released glycans are first permethylated to increase sensitivity of detection and then mixed with suitable matrix before being spotted onto a target plate where ions are generated by laser pulse and separated according to mass by their time of flight (TOF)[25, 40, 47]. With this approach, a structure's accurate mass is determined which can be used to elucidate a likely monosaccharide composition (# hexoses, # deoxyhexoses, #sialic acids and # HexNAc's) along with relative glycan abundances. In addition, the few glycan data analysis software tools that have been developed (such as Cartoonist [48] and Glycoworkbench [49]) have been tailored to data generated by this type of approach. Rapid general glycan profile comparisons between samples can be generated to identify possible differences between experimental and control populations of glycans; However, no fragmentation data is obtained. Therefore, structural isomers cannot be differentiated nor can linkages or branching patterns.

The MS^n approach can be divided into two subclasses, MS/MS and MS^n. The MS/MS strategy takes one full MS scan by MALDI/TOF to identify potential glycan molecular ions and then subjects these ions to one degree of fragmentation by MALDI-TOF/TOF CID looking for corroborating fragment ions to support the proposed glycan composition[39]. Occasionally, some information regarding sequence and branching information may be obtained as well[50]. The major drawback of adding the step of fragmentation is the exponential increase in the time and skill level required for data interpretation.

The MS^n approach employs at least 3 rounds of fragmentation and is the most data rich MS technique employed today; a wealth of structural details can be elucidated by this methodology[51],[52],[53]. In some cases, general composition, sequence and branching points, can be determined by MS^n. For example, the location of a fucose (core or distal, as well as linear or branched as in Le^x) on a complex N-glycan can be determined by the presence and absence of key fragment ions. The MS^n approach involves *de novo* sequencing of each glycan present and requires expertise as well as a substantial time investment for data interpretation, especially considering the lack of software currently available. Detection of differences in these structural details however, provides the opportunity to observe significant changes in the fine details of glycosylation that would otherwise not be seen.

Many variations and combinations of the basic MS analytical strategies exist. ESI ionization is commonly combined with a variety of ion trap mass analysis techniques including linear ion trap (LTQ), Orbitrap, and FT-ICR, depending on availability and the type of information desired. A plethora of instrument workflows within each variation also exists a few of which include total ion mapping (TIM), selected reaction monitoring (SRM)[54], and selected ion monitoring (SIM)[55] analysis. TIM analysis, involves successive isolation and fragmentation of overlapping ranges of ions effectively creating a MS/MS map of all ions present. SIM and SRM analyses are data dependent acquisition techniques which monitor single product ions or reactions throughout sample injection usually following online chromatographic separation. Separation of native glycans into neutral and acidic pools followed by MALDI-FT-ICR-MS^2 analysis[56] as well as separation of permethylated glycans with a porous graphite column online with a Q- TOF CID produces fragmentation up to MS^3 [57]. Many other combinations exist and new ones are continually being developed.

Mass Spectrometry Data Interpretation

The task of mass spectral data interpretation for glycan analysis is certainly in its infancy and given the recent surge of interest in the field, many advances are expected in the near future. Development of software capable of automating data interpretation is a desperately needed resource for the glycomics community. A broad range of bioinformatic methodologies are beginning to emerge[1, 58]. One strategy of automated interpretation employs algorithms that compare experimentally derived spectra against databases of known glycan structures and scores the confidence level and false discovery rate of an assignment much as is done by the SEQUEST algorithm in proteomics[48]. This approach is commonly used with single MS dimensional data such as that obtained by MALDI/TOF analysis annotated by Cartoonist algorithms[48]. The downside of this approach is that only previously reported glycan structures would be identified and novel structures would not be proposed. As mentioned earlier, the difficulty of creating a quality theoretical glycan database with no template from which to derive glycan sequences, limits the ease of proteomic style scoring.

Sample size is limited and can be completely exhausted by one TIM analysis, yielding only MS^2 data. However, the primary bottle neck of this work flow is data interpretation. Manual interpretation of the 800 spectra generated per TIM is time consuming. Additionally, every glycan in a TIM analysis is first characterized and then compared to other samples in order to detect changes. Manual interpretation is impractical for analyzing the larger datasets that are required for biomarker discovery or screening for cellular changes in glycosylation. Many investigators refer to this step as the bottleneck of glycan analysis and have called for developing strategies to alleviate the challenges discussed.

Spectral interpretation is based on the presence and absence of particular fragment ions[51, 53]. Here, a structure is first fragmented and structural details of the individual pieces are assigned by unique cleavages and the location of free hydroxyl groups that were methylated during derivitazation[59]. A confidence can be assigned to each section of the glycan giving linkage anomerocity and sequence information which together with knowledge of biosynthetic pathways can be used to reconstruct the intact glycan.

In order to get maximum data with minimal time and sample, novel MS^n workflows on a Thermo LTQ Orbitrap™ mass spectrometer[45] have been employed. Workflows follow the basic scheme of direct infusion of free/released permethylated glycans into a ThermoFisher Orbitrap Discovery™ mass spectrometer coupled with electrospray ionization (ESI). Analysis of intact glycan structures is performed in the Orbitrap™ for maximum mass accuracy whereas collision induced dissociation (CID) fragmentation and subsequent mass analysis of fragments is performed in the ion trap for maximum sensitivity is used. The most abundant peak in the full-MS spectrum is automatically selected to be fragmented in the ion trap generating an MS^2 spectra from which product ions may be automatically fragmented further based on the presence of signature fragment ions or detection of a preselected neutral loss (NL) giving MS^3 data (Figure #). Product ions may be selected for further analysis if desired. The parent ion is excluded from being selected again for the remainder of the run and the cycle is repeated, ultimately resulting in a list of m/z present in order of intensity, followed by fragmentation data for each peak selected. These methods are designed to be flexible and can be fine-tuned to focus on the type of data desired. For example, the presence of particular epitopes such as Le^x, can be characterized by probing for specific NL. Sensitivity can be increased by dividing the mass range into smaller regions, for example, run the program from 700-1200 m/z and again from 1200-2000 m/z. A parent mass list can also be generated from the Full MS which is then used to select ions for fragmentation, or, to be excluded if low abundance glycans are the focus.

One example of manual data interpretation strategy is as follows: A full MS is taken and the most abundant peak m/z 1134 (z=+2) is selected for fragmentation generating the second spectrum shown (Figure #). In this case a particular fragment ion (m/z 660) was detected which triggered another level of fragmentation. Based on these three spectra, first, the molecular weight of the ion was calculated (m/z 2245) and entered into GlycoMod yielding the basic monosaccharide composition of M_3N_2 + $(Hex)_2(HexNAc)_2(deoxyhexose)_1$. Next, several plausible structures were proposed based on biosynthetic pathways. The placement of the fucose is the main point of variability in this structure. The Fuc could be located on the core GlcNAc or at the non-reducing terminal (linear or branched) as in the Le^x epitope. The MS^3 spectra indicated that it is not core fucosylated by the presence of m/z 660 and not 474. The MS^4 spectrum of m/z 660 showed a terminal but not Le^x location for the fucose. Without these additional fragmentation steps, accurate structural determination would have been unlikely. Therefore, this level of analysis is crucial for a true understanding of glycosylation patterns in a sample. Most analysis performed today is at the MS or MS^2 level but generates less than 100 partially characterized structures. Additional MS^3 or MS^4 experiments are required to resolve isobaric configurations.

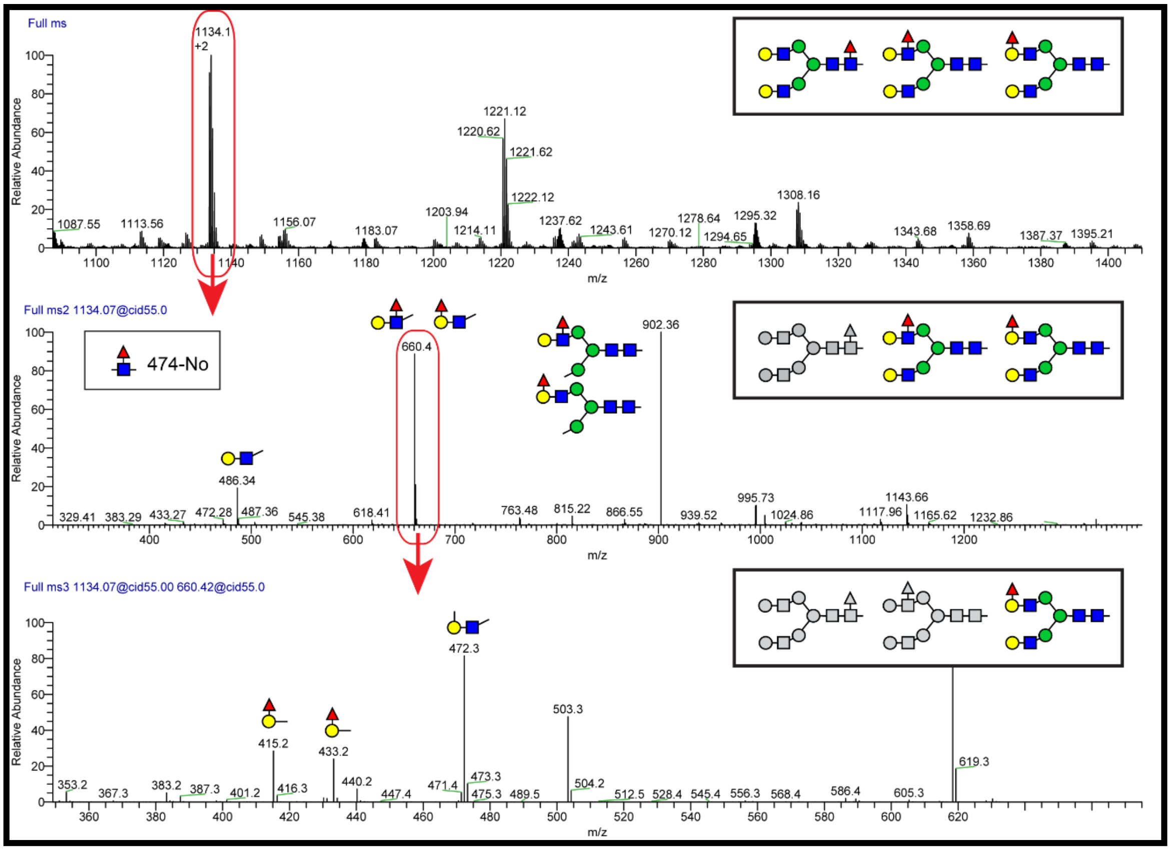


Figure 2.6 Structural determination of glycans MS^3 approach

Bioinformatic Tools For MS Data Interpretation

Bioinformatic tools that are robust, freely available and instrument independent are desperately needed in the glycomics community, especially to assist in the annotation of mass spectrometric analysis of glycans released from glycoproteins. Some tools have been developed over the last decade[51, 60]. However, most are currently unavailable to the community and/or are not supported for further development.

A glycan's basic composition can be determined by MS1 analysis, However, MS/MS is necessary to determine a glycan's sequence or topology. Analysis of MS data cannot determine the type of glycosidic bond (α or β) or distinguish isomeric monosaccharides like glucose/mannose/galactose which have the same mass. It is also difficult to determine which arm of a branched glycan carries specific terminal modifications. Therefore, structural annotations frequently rely on known biosynthetic pathway rules to ascertain these features. For example, it is well established that N-glycans are initiated by a GlcNAc residue attached to an asparagine, therefore assignment of that monosaccharide as GlcNAc rather than its isomer GalNAc is reasonable despite not having MS data that specifically defines that feature.

In addition, most glycan MS analyses rely on permethylation[61] of free hydroxyl to methyl groups to increase sensitivity and improve ionization efficiency for MS analysis by ESI. Permethylation also aids in structural determination because fragmentation of permethylated structures generates scars at sites of substitution. For example, a Hexose-HexNAc disaccharide fragment that was terminally located would have a mass that includes one more methyl group than one that was located internally. This type of information is critical when determining structural variations in glycomics. However permethylation equalizes many of the chemical functionalities that might allow chromatographic separation based on chemical properties. Thus, chromatographic resolution of permethylated N-linked glycans has proven to be difficult to date. Combined with the lack of robust automated software necessary to interpret the thousands of additional spectra, each of which will require manual interpretation, LC-MS or LC-MS/MS for permethylated glycans has been under-developed so far. This shortcoming is unfortunate because LC offers unique opportunities to separate isobaric glycans that are currently unresolved in direct infusion analyses. We currently rely on the presence or absence of diagnostic ions in MS/MS spectra to determine structural topology. However, isobaric mixtures complicate this strategy. Ratios of key diagnostic ions can serve as a comparison between samples to give an indication of changes, but this approach is not reliable for determining the absolute quantity of isoforms in a mixture. The best solution for isomer mixtures will be to develop LC-MS/MS separation strategies and then employ automated annotation tools to interpret the massive volume of spectra that are generated, similar to well-established proteomic approaches.

All glycomic discussions presented here have been focused on released glycans. However, it is desirable to not only know the glycan structure, but also know which site(s) they are attached to and on which proteins. Strategies and tools have begun to emerge for the analyses of intact glycoproteins which usually employ a two pronged approach where users first define the glycome, then using that information, define the peptides with those particular attachments. Software for glycopeptide analysis therefore requires robust tools to first define the glycome. Therefore software tools to determine the released glycans remains a critical first step in the process.

As previously discussed, glycan composition can be determined by MS using intact parent mass without any fragmentation data. GlycoMod[60c], Cooper 2001- attempts to find all possible oligosaccharide compositions that correspond to a particular parent ion mass. It is helpful for determining compositions but does not use MS/MS spectra and therefore cannot differentiate between different topologies. Glycomod is freely available online at http://web.expasy.org/glycomod/ .SysBioWare[62], Vakhrushev 2009, also uses MS1 data to determine glycan composition based on parent ion mass with added features including de-isotoping, baseline adjustment, and denoising. It is not freely available to the public. Cartoonist[63], Goldberg 2006, labels MALDI MS1 glycan data with probable compositions from a database of 300 N-linked structures. CartoonistTwo[60e] captures the composition from Cartoonist and then further considers MS/MS CID data to assign topologies using multiple scoring and calibration approaches. The analysis relies on a set of N-glycans that may or may not exist in nature and neither version is freely accessible to the community.

Additional programs have been developed that utilize SEQUEST like database matching algorithms including Glycoworkbench[64] Ceroni 2008[65] and GlycoFragment[66] GlycoSearchMS[67] Lohmann 2004, which are freely available and SimGlycan©[60a], a commercially available tool Apte 2009. GlycoFragment and GlycoSearchMS are two modules that work together to aid in manual annotations. GlycoSearchMS finds candidate compositions much like GlycoMod[60c] and then utilizes the GlycoFragment module to theoretically fragment candidate structures, thereby producing a list of expected ions for users to manually compare against experimental spectra. Glycoworkbench follows the same strategy but adds the possibility of automatically comparing theoretical peaks with experimental data. These tools are helpful when comparing one spectrum at a time however they are limited when dealing with hundreds and thousands of spectra. SimGlycan© takes these approaches one step further by analyzing spectra in batches of up to 1,500 which improves high throughput capabilities. However, SimGlycan© is proprietary and is limited by cost and access. Additionally, the public databases utilized for these three programs suffer from pollution, including redundancies, omissions and unlikely structures that complicate and slow down analysis time and accuracy.

OSCAR[68] by Ashline 2007[52] uses a complex MSn approach to fully describe and annotate PerMe O-linked glycan structures. The level of structural determination produced by this approach is impressive. However, sample size can be a limiting factor since MS5 or more is often required. Additionally, OSCAR is not currently accessible to the public or supported by any appropriate infrastructure.

Several programs have been developed to annotate tandem glycan MS spectra using a de novo sequencing approach in which structures are computed directly from the spectrum without the help of a glycan database. They include: STAT[69] Gaucher 2000, StrOligo[70] Eithier 2002 (N-link instrument specific MALDI MS/MS data), a heuristic algorithm[71] by Shan 2008, and dynamic programming approach called GlyCH[72] by Tang 2005. None of these are available to the public and have considerable limitations given the computation burden associated with the combinations of monosaccharides available without a database to constrain the search space. They are generally limited to no more than 10 monosaccharides per structure, posing significant problems when considering the usual portfolio of N-linked glycans found in biological materials. The UniCarb-DB project initiated in 2009 as a continuation of the EUROCarbDB is currently focused on Glycan MS data and structural assignment based on fragmentation data. However no tools have been provided to the community as of 2014.

Current glycomics algorithms rely heavily on accurate parent ion mass and which peaks are present in the MS/MS spectra to determine the best candidate structures. Each of the approaches described here relies on some type of scoring mechanism. For example, Glycoworkbench has two scores, a counting score which reports the number of theoretical peaks of a glycan that are present in an experimental spectra, and an intensity score which says what percent of the total intensity present in the spectra is explained by those peaks. Both scores are important; a glycan that has a high counting score but a low intensity score could indicate a poor match based on the peaks matching being of very low intensity and therefore possibly noise. However, a glycan that has a low counting score but a high intensity score could be an equally poor match if it happens to contain a single peak that has high intensity. SimGlycan© uses a proprietary scoring algorithm that considers composition from mass accuracy and branching from specific diagnostic ions and combines them into a single score called the proximity score. Unfortunately, the proprietary nature of the SimGlycan© algorithm precludes a robust validation of the approach. The scoring mechanisms used with these approaches fail to give users a real sense of the validity of their data matches. They can tell you which of the structures that algorithm considered matches the data best but they cannot tell you how likely that it is a real match.

The proteomics community faced similar types of struggles before SEQUEST and MASCOT were fully developed. The proteomics community had to find ways of defining the quality of their fitted data and developed standards for the field including proteomics established false discovery rate (FDR). Glycomics is currently facing the same types of challenges but with the added complexities that are imposed by glycan branching and non-template driven biosynthesis. Protein databases are complete and fully annotated since the genomic templates have been sequenced. Glycans have no possibility to achieve this type of reference database and will suffer from this fact in multiple ways. For example, there will always be the possibility of a glycan structure not being in a database which would pose a problem when using database search strategies. Further, the idea of filling databases with all possible combinations of monosaccharides linked in all possible ways would be particularly troublesome given that there are between 1056 and 27648 possible variations of a simple trimer of monosaccharides depending on which constraints you consider versus 6 possibilities for either nucleic acids or amino acids. Since there is no complete database, there is also no way to make a decoy database to determine FDR from.

Probabilistic scoring mechanisms for peptide MS/MS database matching such as XCorr established by Eng[73] Tabb[74] and MacCoss long before FDR was implemented could be tailored to glycomics data to further improve data matching algorithms. After SEQUEST's matching algorithm finds potential matches for spectra in the database, it applies XCorr (a statistical cross correlation function) to assess the quality of matches. XCorr only depends on the quality of the MS/MS spectrum and its fit to the theoretically generated spectrum and is independent of the database itself. It is important for good matching algorithms to consider multiple factors in calculating goodness of fit including significance as well as correctness of a match.

Quality databases and quality transparent goodness of fit indicators[75] are essential for glycomic tools to be of value to the user and the community as a whole. The glycomic community can leverage the work of pioneers in proteomics which laid important theoretical groundwork that may be applicable to glycomics[76]. Many lessons were learned in the early proteomic years that can be a great value in developing sets of quality annotation standards that can be tailored to the unique nature of glycomic data and can serve as a spring board to push glycomics and eventually glycoproteomics forward.

References

1. Aoki-Kinoshita, K. F., An introduction to bioinformatics for glycomics research. *PLoS Comput Biol* **2008,** *4* (5), e1000075.

2. Apweiler, R.; Hermjakob, H.; Sharon, N., On the Frequency of Protein Glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta* **1999,** *1473* (1), 4 - 8.

3. Brooks, S., Strategies for Analysis of the Glycosylation of Proteins: Current Status and Future Perspectives. *Mol Biotechnol* **2009**.

4. Varki, A.; Baum, L.; Bellis, S.; Cummings, R.; Esko, J.; Hart, G.; Linhardt, R.; Lowe, J.; McEver, R.; Srivastava, A.; Sarkar, R., Working group report: the roles of glycans in hemostasis, inflammation and vascular biology. *Glycobiology* **2008,** *18* (10), 747-9.

5. Rudd, P.; Elliott, T.; Cresswell, P.; Wilson, I.; Dwek, R., Glycosylation and the immune system. *Science* **2001,** *291* (5512), 2370-6.

6. Kolarich, D.; Lepenies, B.; Seeberger, P. H., Glycomics, glycoproteomics and the immune system. *Current opinion in chemical biology* **2012,** *16* (1), 214-220.

7. Stevens, J.; Blixt, O.; Tumpey, T. M.; Taubenberger, J. K.; Paulson, J. C.; Wilson, I. A., Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *science* **2006,** *312* (5772), 404-410.

8. (a) Roche Laboratories Inc. 2009; (b) GlaxoSmithKline. 2009.

9. Stevens, J.; Blixt, O.; Glaser, L.; Taubenberger, J. K.; Palese, P.; Paulson, J. C.; Wilson, I. A., Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *Journal of molecular biology* **2006,** *355* (5), 1143-1155.

10. Kwong, P. D.; Wyatt, R.; Robinson, J.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A., Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **1998,** *393* (6686), 648-659.

11. Walker, L. M.; Phogat, S. K.; Chan-Hui, P.-Y.; Wagner, D.; Phung, P.; Goss, J. L.; Wrin, T.; Simek, M. D.; Fling, S.; Mitcham, J. L., Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* **2009,** *326* (5950), 285-289.

12. Karlyshev, A. V.; Ketley, J. M.; Wren, B. W., The Campylobacter jejuni glycome. *FEMS microbiology reviews* **2005,** *29* (2), 377-390.

13. Varki, A.; Varki, N. M.; Borsig, L., Molecular basis of metastasis. *N Engl J Med* **2009,** *360* (16), 1678-9; author reply 1679-80.

14. Marth, J.; Grewal, P., Mammalian glycosylation in immunity. *Nat Rev Immunol* **2008,** *8* (11), 874-87.

15. Varki, A., *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y., 2009.

16. Consortium for Functional Glycomics. http://www.functionalglycomics.org.

17. Moremen, K. W.; Tiemeyer, M.; Nairn, A. V., Vertebrate protein glycosylation: diversity, synthesis and function. *Nature Reviews Molecular Cell Biology* **2012,** *13* (7), 448-462.

18. Domon, B.; Costello, C. E., Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry. *Biochemistry* **1988,** *27* (5), 1534-43.

19. Varki, N.; Varki, A., Diversity in cell surface sialic acid presentations: implications for biology and disease. *Lab Invest* **2007,** *87* (9), 851-7.

20. Dwek, R., Glycobiology: Toward Understanding the Function of Sugars. *Chem Rev* **1996,** *96* (2), 683-720.

21. (a) Brooks, S., Appropriate glycosylation of recombinant proteins for human use: implications of choice of expression system. *Mol Biotechnol* **2004,** *28* (3), 241-55; (b) Brooks, S., Protein glycosylation in diverse cell systems: implications for modification and analysis of recombinant proteins. *Expert Rev Proteomics* **2006,** *3* (3), 345-59.

22. Varki, A.; Freeze, H.; Manzi, A., Overview of glycoconjugate analysis. *Curr Protoc Protein Sci* **2001,** *Chapter 12,* Unit 12.1.

23. Patel, T.; Parekh, R., Release of oligosaccharides from glycoproteins by hydrazinolysis. *Methods Enzymol* **1994,** *230,* 57-66.

24. Morelle, W.; Guyétant, R.; Strecker, G., Structural analysis of oligosaccharide-alditols released by reductive beta-elimination from oviducal mucins of Rana dalmatina. *Carbohydr Res* **1998,** *306* (3), 435-43.

25. Morelle, W.; Faid, V.; Chirat, F.; Michalski, J. C., Analysis of N- and O-linked glycans from glycoproteins using MALDI-TOF mass spectrometry. *Methods Mol Biol* **2009,** *534*, 5-21.

26. (a) Hagglund, P.; Bunkenborg, J.; Elortza, F.; Jensen, O. N.; Roepstorff, P., A New Strategy for Identification of N-Glycosylated Proteins and Unambiguous Assignment of Their Glycosylation Sites Using HILIC Enrichment and Partial Deglycosylation. *Journal of Proteome Research* **2004,** *3* (3), 556-566; (b) Morelle, W.; Michalski, J. C., Analysis of protein glycosylation by mass spectrometry. *Nat Protoc* **2007,** *2* (7), 1585-602.

27. Rohrer, J., Analyzing sialic acids using high-performance anion-exchange chromatography with pulsed amperometric detection. *Anal Biochem* **2000,** *283* (1), 3-9.

28. Harvey, D. J., Collision-induced fragmentation of negative ions from N-linked glycans derivatized with 2-aminobenzoic acid. *J Mass Spectrom* **2005,** *40* (5), 642-53.

29. Brooks, S., Strategies for Analysis of the Glycosylation of Proteins: Current Status and Future Perspectives. *Molecular Biotechnology*.

30. Jackson, P., Polyacrylamide gel electrophoresis of reducing saccharides labeled with the fluorophore 2-aminoacridone: subpicomolar detection using an imaging system based on a cooled charge-coupled device. *Anal Biochem* **1991,** *196* (2), 238-44.

31. Sharon, N.; Lis, H., History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology* **2004,** *14* (11), 53R-62R.

32. Rudd, P. M.; Dwek, R. A., Rapid, sensitive sequencing of oligosaccharides from glycoproteins. *Current Opinion in Biotechnology* **1997,** *8* (4), 488-497.

33. Yagi, H.; Kato, K., Multidimensional HPLC mapping method for the structural analysis of anionic N-glycans. *Trends in Glycoscience and Glycotechnology* **2009,** *21* (118), 95-104.

34. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** *422* (6928), 198-207.

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

36. (a) Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry* **1991,** *63* (24), 1193A-1203A; (b) Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* **1988,** *60* (20), 2299-2301.

37. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization–principles and practice. *Mass Spectrometry Reviews* **1990,** *9* (1), 37-70.
38. Cole, R. B., Electrospray ionization mass spectrometry: fundamentals, instrumentation, and applications. **1997**.

39. Morelle, W.; Slomianny, M. C.; Diemer, H.; Schaeffer, C.; van Dorsselaer, A.; Michalski, J. C., Fragmentation characteristics of permethylated oligosaccharides using a matrix-assisted laser desorption/ionization two-stage time-of-flight (TOF/TOF) tandem mass spectrometer. *Rapid Commun Mass Spectrom* **2004,** *18* (22), 2637-49.

40. Harvey, D. J., Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update for 2003-2004. *Mass Spectrom Rev* **2009,** *28* (2), 273-361.

41. Finnigan, R. E., Quadrupole mass spectrometers. *Analytical Chemistry* **1994,** *66* (19), 969A-975A.

42. Yost, R. A.; Boyd, R. K., [7] Tandem mass spectrometry: Quadrupole and hybrid instruments. *Methods in enzymology* **1990,** *193*, 154-200.

43. Easterling, M. L.; Mize, T. H.; Amster, I. J., Routine part-per-million mass accuracy for high-mass ions: Space-charge effects in MALDI FT-ICR. *Analytical chemistry* **1999,** *71* (3), 624-632.

44. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R., The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry* **2005,** *40* (4), 430-443.

45. Scientific, T. http://www.thermo.com.

46. Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S., Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry* **2006,** *78* (7), 2113-2120.

47. Harvey, D. J., Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update covering the period 1999-2000. *Mass Spectrom Rev* **2006,** *25* (4), 595-662.

48. Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* **2005,** *5* (4), 865-75.

49. Ceroni, A.; Dell, A.; Haslam, S. M., The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* **2007,** *2*, 3.

50. Goldberg, D.; Bern, M.; North, S.; Haslam, S.; Dell, A., Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics* **2009,** *25* (3), 365-71.

51. Ashline, D.; Singh, S.; Hanneman, A.; Reinhold, V., Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MSn. *Anal Chem* **2005,** *77* (19), 6250-62.

52. Ashline, D. J.; Lapadula, A. J.; Liu, Y. H.; Lin, M.; Grace, M.; Pramanik, B.; Reinhold, V. N., Carbohydrate structural isomers analyzed by sequential mass spectrometry. *Anal Chem* **2007,** *79* (10), 3830-42.

53. Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal Chem* **2005,** *77* (19), 6271-9.

54. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **2008,** *4*, 222.

55. Babu, P.; North, S. J.; Jang-Lee, J.; Chalabi, S.; Mackerness, K.; Stowell, S. R.; Cummings, R. D.; Rankin, S.; Dell, A.; Haslam, S. M., Structural characterisation of neutrophil glycans by ultra sensitive mass spectrometric glycomics methodology. *Glycoconj J* **2008**.

56. Lebrilla, C.; An, H., The prospects of glycan biomarkers for the diagnosis of diseases. *Mol Biosyst* **2009,** *5* (1), 17-20.

57. Costello, C.; Contado-Miller, J.; Cipollo, J., A glycomics platform for the analysis of permethylated oligosaccharide alditols. *J Am Soc Mass Spectrom* **2007,** *18* (10), 1799-812.

58. (a) von der Lieth, C. W.; Bohne-Lang, A.; Lohmann, K. K.; Frank, M., Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform* **2004,** *5* (2), 164-78; (b) von der Lieth, C.; Lütteke, T.; Frank, M., The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim Biophys Acta* **2006,** *1760* (4), 568-77.

59. Zhang, H.; Singh, S.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Anal Chem* **2005,** *77* (19), 6263-70.

60. (a) Apte, A.; Meitei, N. S., Bioinformatics in glycomics: Glycan characterization with mass spectrometric data using SimGlycan™. *Functional Glycomics* **2010**, 269-281; (b) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans†. *Journal of proteome research* **2008,** *7* (4), 1650-1659; (c) Cooper, C.; Gasteiger, E.; Packer, N., GlycoMod--a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **2001,** *1* (2), 340-9; (d) Ethier,

M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H., Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid communications in mass spectrometry* **2003,** *17* (24), 2713-2720; (e) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A., Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res* **2007,** *6* (10), 3995--4005.

61. Aoki, K.; Perlman, M.; Lim, J. M.; Cantu, R.; Wells, L.; Tiemeyer, M., Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo. *Journal of Biological Chemistry* **2007,** *282* (12), 9127-9142.

62. Vakhrushev, S. Y.; Dadimov, D.; Peter-KataliniÄ‡, J., Software platform for high-throughput glycomics. *Anal Chem* **2009,** *81* (9), 3252--3260.

63. Goldberg, D.; Bern, M.; Li, B.; Lebrilla, C., Automatic determination of O-glycan structure from fragmentation spectra. *J Proteome Res* **2006,** *5* (6), 1429-34.

64. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* **2008,** *7* (4), 1650-9.

65. Ceroni, A.; Dell, A.; Haslam, S., The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* **2007,** *2*, 3.

66. Joshi, H. J.; Harrison, M. J.; Schulz, B. L.; Cooper, C. A.; Packer, N. H.; Karlsson, N. G., Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* **2004,** *4* (6), 1650--1664.

67. Lohmann, K. K.; von der Lieth, C.-W., GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res* **2004,** *32* (Web Server issue), W261--W266.

68. Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal Chem* **2005,** *77* (19), 6271--6279.

69. Gaucher, S. P.; Morrow, J.; Leary, J. A., STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Analytical chemistry* **2000,** *72* (11), 2331-2336.

70. Ethier, M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H., Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003,** *17* (24), 2713--2720.

71. Shan, B.; Ma, B.; Zhang, K.; Lajoie, G., Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *Journal of bioinformatics and computational biology* **2008,** *6* (01), 77-91.

72. Tang, H.; Mechref, Y.; Novotny, M. V., Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* **2005,** *21 Suppl 1*, i431--i439.

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

74. Tabb, D. L.; Eng, J. K.; Yates Iii, J. R., Protein identification by SEQUEST. In *Proteome Research: Mass Spectrometry*, Springer: 2001; pp 125-142.

75. MacCoss, M. J.; Wu, C. C.; Yates, J. R., Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical chemistry* **2002,** *74* (21), 5593-5599.

76. Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical chemistry* **2003,** *75* (10), 2470-2477.

CHAPTER 3

DISCRIMINATION BETWEEN ADENOCARCINOMA AND NORMAL PANCREATIC DUCTAL FLUID BY PROTEOMIC AND GLYCOMIC ANALYSIS[1]

[1] 2014, *Journal of Proteome Research*, 13 (2), 395-407, Reprinted here with permission of American Chemical Society Copyright 2014

Mindy Porterfield, Peng Zhao, Haiyong Han, John Cunningham, Kazuhiro Aoki, Daniel D. Von Hoff, Michael J. Demeure, J. Michael Pierce, Michael Tiemeyer, and Lance Wells

Abstract

Sensitive and specific biomarkers for pancreatic cancer are currently unavailable. The high mortality associated with adenocarcinoma of the pancreatic epithelium justifies the broadest possible search for new biomarkers that can facilitate early detection or monitor treatment efficacy. Protein glycosylation is altered in many cancers, leading many to propose that glycoproteomic changes may provide suitable biomarkers. In order to assess this possibility for pancreatic cancer, we have performed an in-depth LC-MS/MS analysis of the proteome and MS^n-based characterization of the N-linked glycome of a small set of pancreatic ductal fluid obtained from normal, pancreatitis, intraductal papillary mucinous neoplasm (IPMN), and pancreatic adenocarcinoma patients. Our results identify a set of seven proteins that were consistently increased in cancer ductal fluid compared to normal (AMYP, PRSS1, GP2-1, CCDC132, REG1A, REG1B, and REG3A) and one protein that was consistently decreased (LIPR2). These proteins are all directly or indirectly associated with the secretory pathway in normal pancreatic cells. Validation of these changes in abundance by Western blotting revealed increased REG protein glycoform diversity in cancer. Characterization of the total N-linked glycome of normal, IPMN, and adenocarcinoma ductal fluid clustered samples into three discrete groups based on the prevalence of 6 dominant glycans. Within each group, the profiles of less prevalent glycans were able to distinguish normal from cancer on this small set of samples. Our results emphasize that individual variation in protein glycosylation must be considered when assessing the value of a glycoproteomic marker, but also indicate that glycosylation diversity across human subjects can be reduced to simpler clusters of individuals whose N-linked glycans share structural features.

Introduction

In 2010, the American Cancer Society estimated 41,000 diagnoses of pancreatic cancer in the U.S. [1]. With a very low percentage of five-year survival, early stage biomarkers for this disease are urgently needed, although there are markers that are used to monitor the course of disease; e.g., the glycan-specific serum marker, CA19-9 [2, 3]. Proteomic analyses of serum samples from patients with pancreatic ductal adenocarcinoma have yielded important information for developing potential biomarkers [4]. Recent data has demonstrated that pancreatic cancer cells are not always of ductal epithelial origin, but may in fact more frequently arise from acinar cells[5], the primary secretory cell responsible for producing the proteins of pancreatic ductal fluid. Therefore, pancreatic ductal fluid, which is likely to contain proteins released from pancreatic adenocarcinoma, has also been subjected to proteomic analysis in search of markers that could also be present in sera [6-12]. These previous studies have produced proteomes that overlap with each other and with the results reported here. However, full validation of a single proteomic marker or set of proteomic markers has not been achieved for pancreatic cancer.

In addition to altered protein expression during oncogenesis, many studies have clearly documented that the glycans expressed on glycoproteins secreted or released from various types of cancer cells exhibit changes in structure that are cell-type specific. For example, glycoproteins that express N-glycans with a “core fucose” residue (alpha1,6 fucose) are secreted into serum from hepatocellular carcinoma (HCC) but not from cirrhotic hepatocytes. An assay for core-fucosylated alpha-fetoprotein is in use to test for HCC, and there is evidence that including other core-fucosylated glycoproteins, such as

GP73, in the analysis yields an HCC diagnostic test with higher specificity and sensitivity[13-16]. Aberrant glycosylation in pancreatic carcinoma is apparent by increased serum levels of CA19-9 and by the detection of circulating antibodies directed against the mucin MUC1 that expresses truncated O-linked glycans (Tn antigens)[17].

In order to identify additional serum markers for pancreatic carcinoma, particularly those for early detection, our approach has first focused on applying proteomic and glycomic analytical technologies for in-depth analysis of pancreatic ductal fluid. Markers identified in ductal fluid are then candidates for validation as serum markers using antibodies that recognize the glycan and protein differences that are identified between ductal fluid samples from patients with pancreatic cancer and from controls, including pancreatitis and intraductal papillary mucinous neoplasms (IPMN). Here, we report in-depth analysis of the proteome and the N-linked glycome of a training set of ductal fluid samples. The results provide potential targets for full validation and highlight important considerations for analyzing human glycoproteomes.

Experimental Methods

Relative Contributions of Authors

The research presented here was contributed to equally between Peng Zhao under the direction of Lance Wells and Melody Perlman Porterfield under the direction of Michael Tiemeyer. Zhao performed all protein based analyses and Porterfield performed all glycan based analyses.

Pancreatic Ductal Fluid Samples

Pancreatic ductal fluid samples along with matching serum and plasma samples were collected from patients who underwent endoscopic retrograde cholangiopancreatogram (ERCP) or endoscopic ultrasound (EUS) procedures. The ductal fluid samples were snap frozen in liquid nitrogen following aspiration from the patients. Sample collection protocols were reviewed and approved by the Institutional Review Boards (IRBs) at the University of Arizona (Tucson, AZ) and the Translational Genomics Research Institute (Phoenix, AZ), and written informed consent was obtained from all patients. Pancreatic ductal fluid samples from patients with the following four diagnoses were used in this study: pancreatic cancer, intraductal papillary mucinous neoplasm (IPMN), pancreatitis, and normal pancreas (Sphincter of Oddi Dysfunction, SOD Type II or III). The age ranges and gender distributions for each diagnostic class were as follows: pancreatic cancer, 54-79 years, 5 male, 4 female; IPMN, 72-77 years, 2 male, 1 female; pancreatitis, 41-56 years, 1 male, 2 female; normal pancreas, 31-72 years, 2 male, 7 female (Supplementary Table S1). For protein identification and quantification, 12 samples in total were analyzed (three samples for each diagnosis). Clear (no visible blood or bile contamination) pancreatic ductal fluid samples were thawed on ice and filtered by 0.2 μm spin columns (Nanosep). Protein concentration of all the samples was determined using the micro BCA protein assay kit (Pierce) following the manufacturer's instructions. Equal amounts of protein (1 mg) were used for analysis. For glycomic analysis, 3 normal, 4 pancreatic cancer, and 2 IPMN samples were chosen. As described for proteomic analysis, ductal fluid samples that were clear (no visible blood or bile contamination) were selected for glycomic analysis. For each sample, 300-500 μl of ductal fluid were extracted with organic solvent as previously described[18]. Briefly, for 300 μl of ductal fluid, total volume was adjusted to 4:8:3 (chloroform:methanol:water) by the addition of 0.4 ml water, 1.87 ml methanol, and 0.93

ml chloroform. The adjusted sample was extracted overnight by nutation. The next morning, proteins were harvested by centrifugation. The protein pellet was washed three times with cold 80% acetone in water and the final pellet was dried under a stream of nitrogen. The resulting solids were harvested as a uniform, white protein powder, which was stored desiccated at –20 ^{o}C. The amount of material harvested from any individual sample was insufficient to allow glycomic and proteomic analysis to be performed on the same samples.

Trypsin Digestion for Proteomic Analysis

The fluid samples were reduced with 10 mM DTT for 1 h at 56 °C, alkylated (carboxyamidomethylated) with 55 mM iodoacetamide (Sigma) in dark for 45 min, and digested with trypsin (Promega) in 40 mM NH_4HCO_3 overnight at 37 °C. The digestion was quenched with 1% trifluoroacetic acid (TFA), and the resulting peptides were desalted with C18 spin columns (Vydac Silica C18, The Nest Group, Inc.) and dried in a Speed Vac.

Protein Fractionation

Protein fractionation was performed by reverse phase liquid chromatography (RP-LC) using the Agilent 1100 series HPLC system (Agilent Technologies). Solvent A (0.1% TFA) and solvent B (0.085% TFA/80% acetonitrile) were used to develop a linear gradient starting with 5 minutes at 5% solvent B (95% solvent A), followed by a 60-minute gradient at variable slope from 5% to 95% solvent B and staying for 3 minutes at 95% solvent B, then returning to 5% solvent B (95% solvent A) in 1.5 minutes and staying for 4.5 minutes at 5% solvent B (95% solvent A). Dried peptides were dissolved in solvent A and separated on a 2.1 × 250 mm silica-based C18 column (VYDAC) at a flow rate of 100 µl/min over the linear gradient. Eluted peptides were collected every 4 min, and subsequently combined into 5 fractions (F1, 15-32%; F2, 32-40%; F3, 40-45%; F4, 45-55%; and F5, 55-85%), desalted and dried as described above.

Reverse Phase nanoLC-MS/MS Analysis

Dried peptides from each fraction generated by RP-LC (12 x 5 in total) were resuspended in 0.5 µl of solvent B (0.1% formic acid/80% acetonenitrile) and 19.5 µl of solvent A (0.1% formic acid) and loaded onto a 75 µm x 105 mm C18 reverse phase column (packed in house, YMC GEL ODS-AQ120ǺS-5, Waters) by nitrogen bomb. Peptides were eluted directly into the nanospray source of an LTQ Orbitrap XL™ (Thermo Fisher Scientific) with a 140-min linear gradient consisting of 5-100% solvent B over 90-95 min at a flow rate of ~250 nl/min. In order to optimize the separation of peptides eluted into the mass spectrometer, gradients were expanded over a 70-min period in the appropriate region corresponding to each fraction collected from the previous offline RP-LC separation (F1, 4-30%; F2, 9-35%; F3, 15-42%; F4, 20-55%; and F5, 28-85%). The spray voltage was set to 2.0 kV and the temperature of the heated capillary was set to 200 °C. Full scan MS spectra were acquired from m/z 300 to 2000 with a resolution of 60000 at m/z 400 after accumulation of 1000000 ions (mass accuracy < 5 ppm). MS/MS events were triggered by the 6 most intense ions from the preview of full scan and a dynamic exclusion window was applied which prevents the same m/z value from being selected for 6 seconds after its acquisition. All 5 sub-fractions were analyzed in technical triplicates and data were acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific). Spectra will be made available upon request.

Proteomic Data Analysis

The acquired MS/MS spectra were searched against the UniProt human proteome database (58831 entries, updated at May 10, 2009) using SEQUEST (Bioworks 3.3, Thermo Fisher Scientific) with the following settings: 50-ppm and 0.5-Da deviation were set for monoisotopic precursor and fragment masses, respectively; trypsin was specified as the enzyme; only fully tryptic peptide identifications were retained; a maximum of 3 missed cleavage sites, 3 differential amino acids per modification and 3 differential modifications per peptide were allowed; oxidized methionine (+15.9949 Da) and carbamidomethylated cysteine (+57.0215 Da) were set as differential modifications. All of the raw spectra were searched against both normal and reversed database under the same parameters, and all of the output files from SEQUEST search were filtered and grouped by different biological samples and replicates in ProteoIQ™. The cutoff value of peptides was set to an Xcorr of 0.5 and the minimum peptide length was set to 4 amino acids. For protein identification, false discovery rate was set to 1% at protein level and peptides matched to multiple proteins were excluded; for protein quantification, the 1% protein-level false discovery rate data was further filtered to achieve a 10% peptide-level false discovery rate, and only proteins that are identified by more than one peptide and in more than one biological sample were considered. The validated result was submitted to Gene Ontology (www.geneontology.org)[19] for protein subcellular localization and biological function annotation.

In order to compare the protein expression levels across samples with different diagnoses, normalized spectral abundance factors (NSAF) were calculated for each protein that was commonly observed in all four diagnoses. In this approach, the spectral counts (SpC) of each protein in a given dataset were divided by its length (L) and normalized to the sum of SpC/L values in the given dataset[20, 21]:

$$NSAF_x = \frac{(\frac{SpC}{L})_x}{\sum_{i=1}^{N}(\frac{SpC}{L})_x}$$

To further resolve shared peptides between protein isoforms, a distribution factor was introduced into the calculation of NSAF[22]:

$$dNSAF = \frac{uSpC + [(d)(sSpC)]}{uL + sL}$$

$$d = \frac{uSpC}{\sum uSpC}$$

According to the equations above, dNSAF is calculated where spectral counts from shared peptides are distributed among protein isoforms based on a distribution factor, d. Spectral counts from peptides uniquely mapping to a protein are denoted as “uSpC”, while spectral counts from peptides shared between isoforms are labeled “sSpC”. Protein amino acid lengths mapped to unique and shared peptides are denoted as “uL” and “sL”, respectively.

Immunoblotting

Protein concentrations of normal and cancer pancreatic juice samples were determined by micro BCA protein assay. Equal amounts of protein from normal and cancer samples (ranging from 2-8 µg for different antibodies) were separated by 4-20% Tris-HCl precast minigels (Bio-Rad), and semi-dry transferred to Immobilon-P transfer membrane (Millipore). The membranes were blocked with 5% BSA in TBST (TBS with 0.1% Tween 20), and probed with each antibody at 4 °C overnight as follows: 1:1000 dilution for REG1A (Abcam), REG1B (Abcam), and REG3A (Abnova) blots, and 1:2000 dilution for phospholipase A2 (Abcam) and pancreatic lipase-related protein 2 (Abnova) blots. After the addition of secondary antibodies conjugated to horseradish peroxidase (HRP) at room temperature for 1 h, the final detection of HRP activity was performed using SuperSignal West Pico chemiluminescent substrates (Thermo Fisher Scientific). The films were exposed to CL-XPosure film (Thermo Fisher Scientific). The amount of material harvested from any individual sample was insufficient to allow glycomic and proteomic analysis to be performed on the same samples., and orthogonal analyses (western blotting) to all be performed on the same samples. However, in some cases, proteomic and western blot anlaysis were performed on the same sample.

N-linked Glycan Analysis

N-linked glycans were prepared from tryptic/chymotryptic digests of total ductal fluid proteins as described previously[18]. Briefly, protein powder produced by organic extraction of ductal fluid (described above) was resuspended in 200 μl of trypsin buffer (0.1 M Tris-HCl, pH 8.2 containing 1 mM $CaCl_2$) by sonication and boiling for 5 minutes. After cooling to room temperature, 25 μl of trypsin solution (2 mg/ml in trypsin buffer) and 25 μl of chymotrypsin solution (2mg/ml in trypsin buffer) were added. Digestion was allowed to proceed for 18 hours at 37 ^{o}C before the mixture was boiled for 5 minutes. Insoluble material was removed by centrifugation and the supernatant was removed and dried by vacuum centrifugation. The dried peptide mixture was resuspended in 250 μl of 5% (v/v) acetic acid and loaded onto a Sep-Pak C18 cartridge column. The cartridge was washed with 10 column volumes of 5% acetic acid. Glycopeptides were then eluted step-wise, first with 3 volumes of 20% isopropanol in 5% acetic acid and then with 3 volumes of 40% isopropanol in 5% acetic acid. The 20% and

40% isopropanol steps were pooled and evaporated to dryness. Dried glycopeptides were resuspended in 50 μl of 20 mM sodium phosphate buffer, pH 7.5, for digestion with PNGaseF (Prozyme, San Leandro, CA). Following PNGaseF digestion for 18 hr at 37 ^{o}C, released oligosaccharides were separated from peptide and enzyme by passage through a Sep-Pak C18 cartridge. The digestion mixture was adjusted to 5% acetic acid and loaded onto the Sep-Pak. The column run-through and an additional wash with 3 column volumes of 5% acetic acid, containing released oligosaccharides, were collected together and evaporated to dryness.

Following enzymatic release and clean-up, liberated N-linked glycans were permethylated[23] and analyzed by direct infusion, nanospray ionization, ion trap mass spectrometry (NSI-LTQ/Orbitrap, Thermo Fisher). An automated MS workflow was employed to sequentially capture MS/MS spectra for all detectable ions. In this workflow, full MS spectra were obtained in the Orbitrap and the highest intensity peak was then selected for fragmentation in the linear trap (collision energy was 35-55% based on instrument calibration). Following acquisition of each MS/MS spectra, the next most intense parent ion was selected for fragmentation. In order to limit the fragmentation of redundant isotopes, an m/z window extending from -1.2 mass units below to +2.1 mass units above the parent ion was excluded. The cycle was repeated until fragmentation profiles revealed only background noise, generally 200 rounds. The resulting MS and MS/MS files were processed using SimGlycan© (Premier Biosoft, Palo Alto, CA) to provide initial structural assignments for all m/z values associated with glycans[24]. SimGlycan assignments were subsequently validated by manually inspecting MS/MS spectra for the presence of signature fragments consistent with the proposed structure for all glycans that demonstrated signal intensity differences in cancer or IPMN greater than 2-fold above or below normal samples (61 assignments). SimGlycan© assignments were also validated manually if the assigned structure was not considered to be a likely component of the human ductal fluid glycoproteome (Xylose-containing glycans, inappropriately degraded structures, biosynthetic impossibilities). Such artifactual assignments arise because these structures are contained within the database used by SimGlycan©. When MS/MS spectra for such candidate glycans were manually inspected, they uniformly revealed a lack of glycan-based fragmentation and their intensities were excluded from the total profile. Signal intensities for valid glycan assignments were retrieved from full MS spectra as peak areas obtained using the Orbitrap FT. Signals associated with different charge states of the same glycans were combined. The prevalence of each glycan was calculated by normalizing its signal intensity to the total signal intensity for all detected glycans and is expressed as "% Total Profile" for each glycan. The associations of glycans with clinical status were queried by hierarchical clustering methods using Euclidean distance calculations as previously described[25, 26].

Structural assignments for the glycans detected at the reported m/z values were based on the compositions determined by accurate mass of the intact molecule (detected by Orbitrap FT), the presence of diagnostic MS/MS fragments that report specific N-glycan features, and the limitations imposed on structural diversity by known glycan biosynthetic pathways. Key structural features that were used to assign glycan topologies included the detection of B-ion fragments and their Y-ion neutral loss counterparts corresponding to terminal LacNAc (Hex-HexNAc, assumed to be Gal-GlcNAc; fragment at m/z = 486.2, $[m+Na]^+$), internal LacNAc (Hex-HexNAc, fragment at m/z = 472.2, $[m+Na]^+$) sialic acid (fragment at m/z = 398.2, $[m+Na]^+$), outerarm Fuc (as fucosylated LacNAc; deoxyHex-Hex-HexNAc and/or Hex-(deoxyHex)-HexNAc; fragment at m/z = 660.3, $[m+Na]^+$), terminal Fuc (deoxyHex-Hex; fragment at m/z = 415.2, $[m+Na]^+$); core Fuc (as Fuc-HexNAc at the reducing terminal; fragment at m/z = 474.2, $[m+Na]^+$). It is frequently not possible to unambiguously assign non-reducing terminal modifications to a specific arm of a complex N-linked glycan solely using MS/MS spectra. For consistency of presentation and ease of comparison, outer arm modifications are presented as elaborations on the increasingly complex products of the known branching N-acetylglucosamine transferases (GlcNAcT) in the following succession: GlcNAcT1, T2, T4, T5. For example, a monosialylated, fully galactosylated triantennary glycan is depicted with a single sialic acid on the arm initiated by GlcNAcT1 (the 3-arm) and the three antennae would be represented as products of GlcNAcT1, 2, and 4 (two GlcNAc residues on the 3-arm and 1 on the 6-arm). The disialylated form of the same triantennary glycan would be depicted with the second sialic acid added to the arm initiated by GlcNAcT2 (the 6-arm). Structural ambiguity is also annotated by brackets, which are meant to indicate equally likely sites for elongation.

Results

Protein Identification

For each of the 4 diagnoses, 3 patient samples were analyzed in technical triplicates. Each sample was trypsin digested and separated by off-line RP-HPLC separation. Five fractions were collected for each sample and analyzed by LC-MS/MS to yield a total of 45 LC-MS/MS experiments for each diagnosis (180 total LC-MS/MS analyses). After filtering and removing duplicates, the combined data set consists of 368 unique proteins identified by 1995 peptides corresponding to 58930 spectra, 74% (273/368) of which were identified by more than one peptide. Specifically, 112 proteins were identified by 750 peptides with 11598 spectra in normal samples; 138 proteins were identified by 743 peptides with 6590 spectra in pancreatitis samples; 124 proteins were identified by 808 peptides with 22581 spectra in IPMN samples; and 188 proteins were identified by 1068 peptides with 18161 spectra in pancreatic cancer samples (Table 3.1, Table 3.2 and Table 3.3, Figure 3.1). All the identified proteins were submitted to Gene Ontology (www.geneontology.org) for subcellular localization and biological function annotation. Based on the spectral counts assigned to each identified protein, the majority of the proteins are secreted proteins (81%) involved in proteolysis (52%) and metabolic process (29%) (Fig. 3.1).

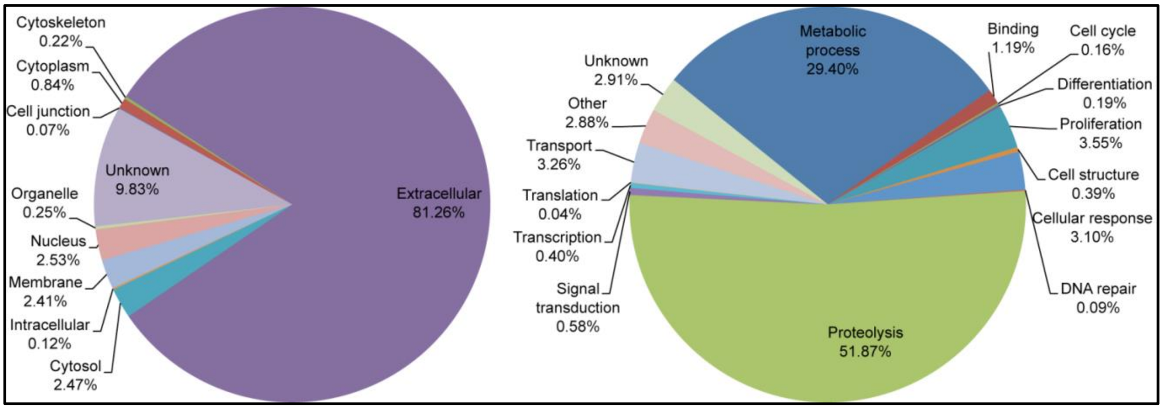


Figure 3.1. Subcellular localization and biological function of proteins identified in 12 pancreatic ductal fluid samples. Distributions were calculated based on spectral counts of identified proteins.

Protein Quantification

To evaluate the variation in protein expression across pancreatic ductal fluid samples with different diagnoses, the identified protein dataset was further filtered to achieve a 10% peptide-level false discovery rate at 1% protein-level false discovery rate. After filtering, the resulting dataset was examined manually to eliminate proteins that were only identified by one peptide or in only one patient. In the final quantified dataset, a total of 47 proteins were quantified with 590 peptides and 46172 spectra across three diagnoses and normal controls (Fig. 3.2). Specifically, 22 proteins were quantified with 300 peptides and 8674 spectra in normal samples; 19 proteins were quantified with 215 and 3774 spectra in pancreatitis samples; 35 proteins were quantified with 414 peptides and 18632 spectra in IPMN samples; and 36 proteins were quantified with 422 peptides and 15092 spectra in pancreatic cancer samples.

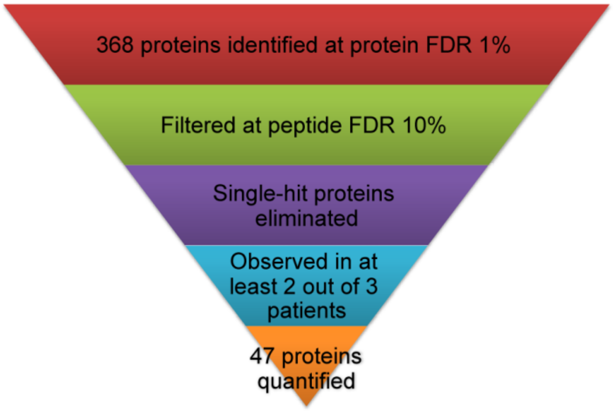

368 proteins identified at protein FDR 1%
Filtered at peptide FDR 10%
Single-hit proteins eliminated
Observed in at least 2 out of 3 patients
47 proteins quantified

Figure 3.2. Data filter process flow chart.
368 proteins were identified by Sequest after filtering at 1% protein-level false-discovery rate (FDR). The dataset was then further filtered at 10% peptide-level FDR and 1-hit proteins were eliminated. In the resulting dataset, only proteins that were observed in at least 2 out of 3 patients were considered for quantification, and finally 47 proteins were quantified.

By comparing the dNSAF values of proteins that were commonly observed in the samples from normal control and the three diagnoses, we were able to discover the differential expression of 22 proteins in our dataset (Fig. 3.3, Table 3.4). In comparison to normal controls, several proteins, such as REG1A, alpha-amylase, trypsin-1, chymotrypsinogen B, and glycoprotein GP2-1, showed significant elevation in IPMN and cancer samples. Several other proteins, such as pancreatic amylase, elastase 2A, 3B and 3A, carboxypeptidase A1, and pancreatic lipase-related protein 2, were downregulated in IPMN and cancer samples compared to normal controls. We also found several proteins that were uniquely expressed in IPMN and/or cancer samples on the quantifiable level (Table 3.5), such as REG1B, REG3A, CCDC132, phospholipase A2, and elastase 2B. As we re-examined the uniquely expressed proteins on the identifiable level, we discovered that even though some of those proteins were unique in IPMN and/or cancer samples on quantifiable level, they may be observed universally in the other biological samples on identifiable level (Table 3.6). For example, REG1B was only seen in two cancer patients on the quantifiable level, however, it was identified in patients with all three diagnoses and normal controls, suggesting the protein is likely present in all samples but upregulated in cancer samples.

Biological variation was also investigated by calculating the standard deviation across the biological triplicates based on the normalized spectral counts of each quantified protein (Figure 3.3, Table 3.7). The pronounced biological variances represented by the data are likely contributed by the differences of individual patients, such as gender, age, blood type and other medical conditions. The statistical data also indicates the need to increase the number of biological samples, and possibly to further stratify the samples based on multiple biological and medical factors instead of solely on diagnosis.

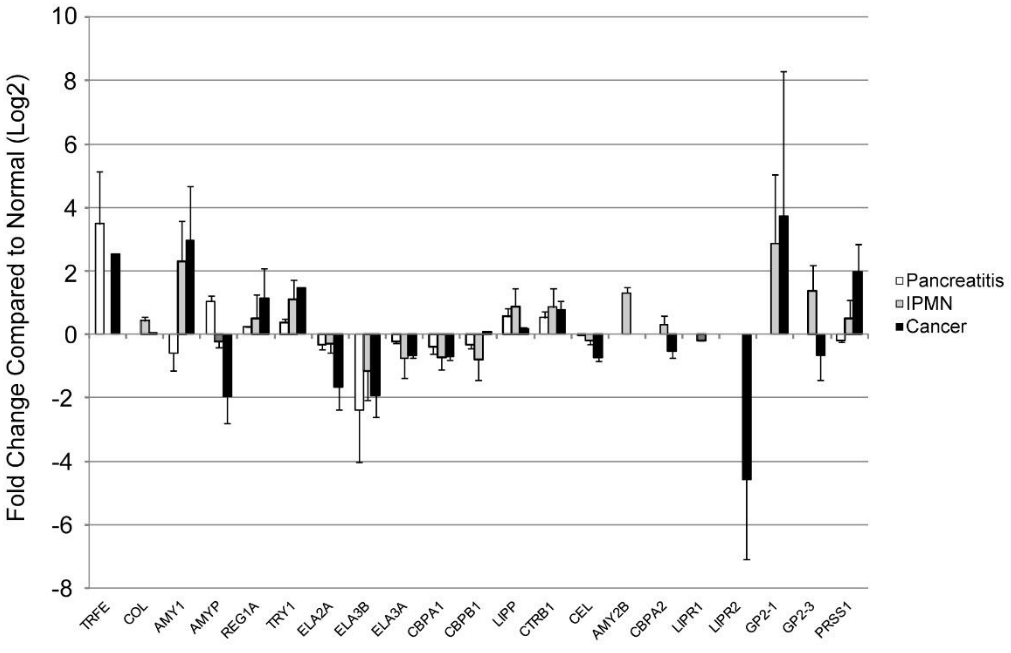


Figure 3.3. Variations in protein expression for pancreatitis, IPMN, and cancer samples. Protein expression variation in pancreatitis, IPMN, and cancer samples in reference to normal controls. The ratios are calculated based on dNASF values of quantified proteins and are plotted on a Log2 scale.

Orthogonal Validation of Protein Identifications

Antibodies were obtained for a subset of candidate biomarkers in order to validate the proteomic results by Western blotting (Fig. 3.4). While normal samples demonstrated 2 major bands for REG1A, additional bands were observed in the cancer sample (Fig. 3.4A). A similar pattern was observed in REG1B and REG3A blots with more prominent increases in abundance and multiple bands present in cancer samples (Fig. 3.4B,C). The molecular weight heterogeneity of REG proteins is believed to result from glycoform heterogeneity and proteolytic processing[27]. Distinctive bands of immunoreactive phospholipase A2 were observed at 32 kDa (full length) and 16 kDa (mature) in the cancer samples and were absent in the normal controls (Fig. 3.4D). Therefore, phospholipase A2 (PLA2) can be considered as a positive marker for pancreatic malignancy. In contrast to REG proteins and PLA2, immunoreactive bands of pancreatic lipase-related protein 2 (LIPR2) at 37 kDa (mature) and at 52 kDa (full length) were decreased in cancer (Fig. 3.4E). Therefore, REG proteins and PLA2 may be positive indicators for pancreatic cancer while LIPR2 may be considered a negative indicator.

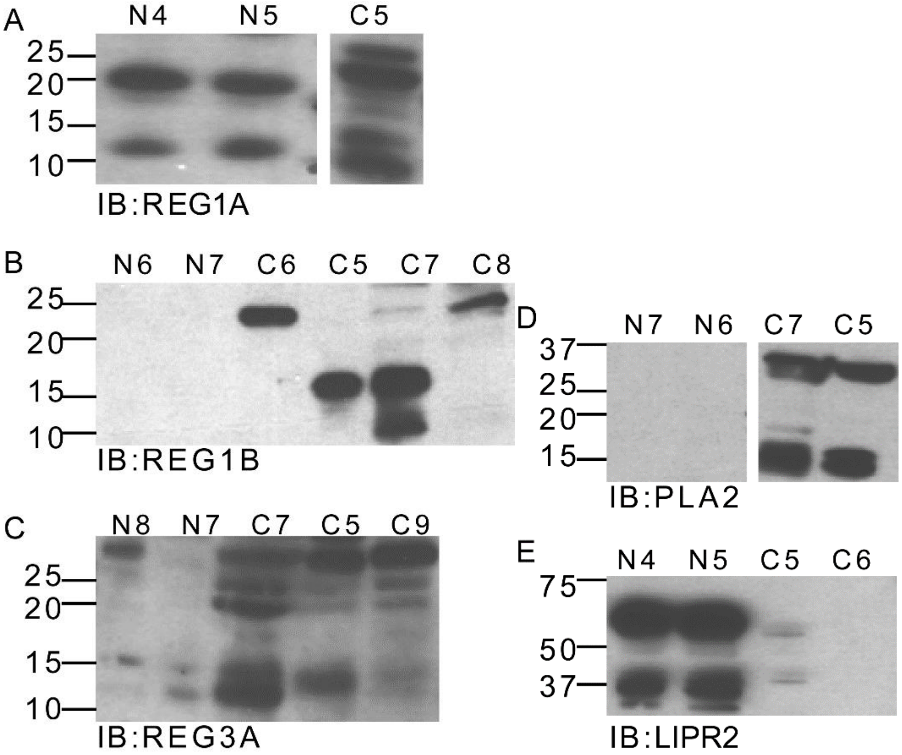


Figure 3.4. Validation of proteomic data by immunoblotting. Pancreatic ductal fluid samples with diagnosis of pancreatic cancer (C5, C6, C7, C8, and C9) were compared to normal controls (N4, N5, N6, N7, and N8) by probing with respective antibodies: (A) REG1A, (B) REG1B, (C) REG3A, (D) Phospholipase A2 (PLA2), and (E) Pancreatic lipase-related protein 2 (LIPR2). Numbers on the left side of the blots indicate molecular weights in kDa. The split panels in A and D were originally part of the same blot, one for A and one for D. The lanes of interest were originally separated by irrelevant samples and have been brought together to facilitate direct comparison.

Total N-linked Glycan Profile

A total of 80 glycans were analyzed by NSI-MS/MS (nanospray ionization-MS/MS) and MS^n as needed to elucidate the structural features of N-linked glycans harvested from 3 normal, 2 IPMN, and 4 pancreatic cancer ductal fluid samples. Comparisons of the prevalence of all N-linked glycans did not detect glycan markers or even glycan patterns that could distinguish cancer from normal (Figure 3.5A). However, the total glycan profiles for the samples analyzed were dominated by a small set of glycans whose prevalences ranged from 6 – 38% of the total profile. These driver glycans overwhelmed the contribution of less prevalent glycans and did not sort with normal, cancer, or IPMN, nor were they correlated with patient gender or age (Table 3.8). After removing the driver glycans from the total profile and recalculating the prevalence of the remaining glycans, differences in the profile of minor glycans became apparent. By Wilcoxon rank-sum test, 9 of the remaining glycans showed increased or decreased prevalence ($p \leq 0.05$) in cancer or IPMN compared to normal (Fig. 3.6). Several of the discriminating glycan structures carry blood group epitopes of the H, Lewis X/A or Lewis Y/B type. However, in their entirety, blood group epitopes or secretor status were not able to sort the samples by diagnosis, indicating that blood group by itself does not account for the observed segregation of cancer, normal, and IPMN (Figure 3.5B).

A striking division of the 9 samples was detected by comparing the prevalences of the driver glycans. All of the analyzed samples could be assigned to 1 of 3 groups based on driver glycan prevalence: S-Group, dominated by sialylated glycans; F-Group, dominated by fucosylated glycans; M-Group, characterized by a mixture of the dominant S and F glycans (Fig. 3.7, 3.8, and 3.9). Within each group, N-linked glycan profiles segregated normal from cancer (Fig. 3.10 and 3.11). For the S-group, cancer samples showed increases in the major driver glycans (structures 34, 50, 55) as well as increases in branching and additional sialylation of less prevalent glycans. For the M-group, an obvious trend was not discernable among the major driver glycans. However, among the less prevalent glycans in the M-group, increases in high mannose (structures 4, 6, 13, 22, 21) and less complex glycans (structures 9, 17) mirrored decreases in highly branched, fully galactosylated and fucosylated glycans in the cancer and IPMN samples. For the F-group, the cancer sample showed decreases in the major driver glycans (structures 26, 33, 40) as well as increases in glycan branching, outer arm fucosylation, and poly-LacNAc extension (structures 75, 79, 81, 82, 85). The glycan structures that define the S-, M-, and F-groups (Fig. 3.7) are not biosynthetic precursors for the glycan structures that differentiate cancer from control samples within each group (Fig. 3.10 and 3.11), indicating that the generation of putative marker glycans does not simply reflect the up-regulation of a dominant glycan processing pathway.

Interestingly, the glycan profiles of the IPMN samples tested here exhibited characteristics of normal and cancer samples. Consequently, the IPMN sample assigned to the F-group clustered with normal and the IPMN sample in the M-group clustered with cancer. The segregation of the IPMN samples likely reflects the transitional nature of this diagnosis, with IPMN patients exhibiting a continuum of clinical presentation, including the possible progression to adenocarcinoma.

Figure 3.5

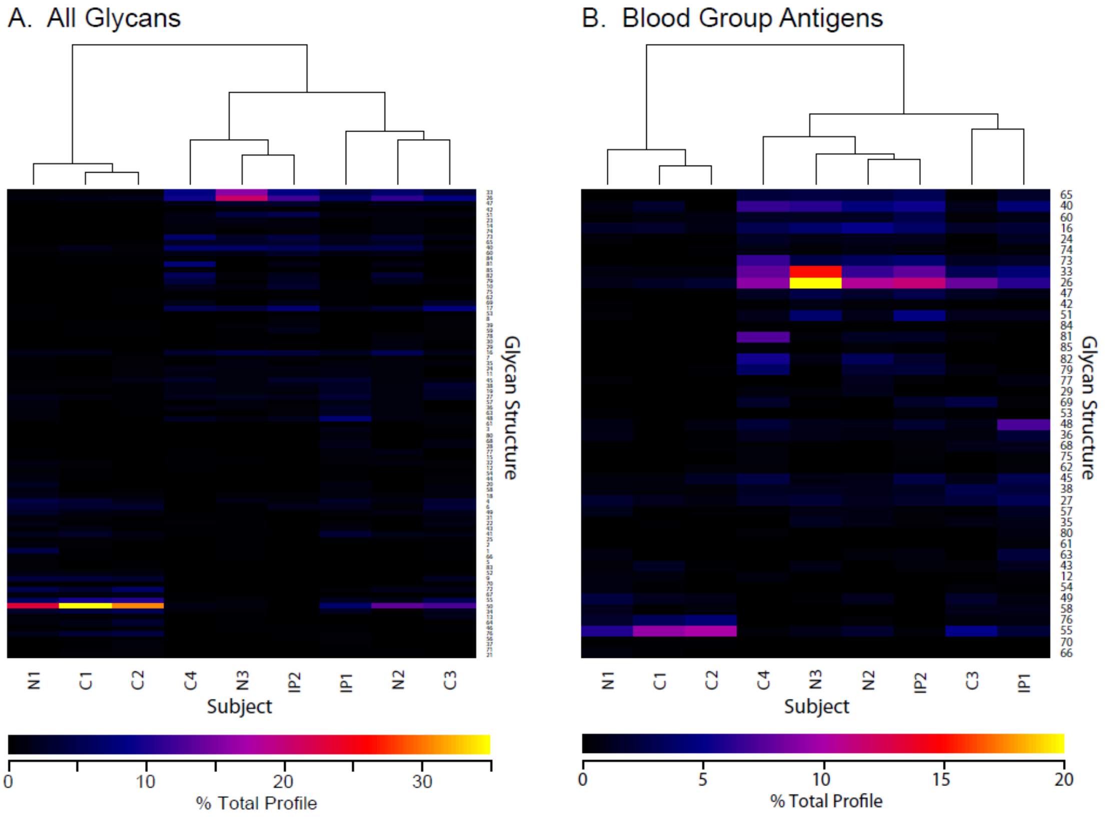


Figure 3.5. Heirarchical clustering of all detected glycans and blood group antigens. (**A**) The prevalence of glycans in the total profile of all detected glycans does not discriminate between cancer, IPMN, and normal.

(**B**) Glycans bearing blood group epitopes also do not discriminate between cancer, IPMN, and normal samples. Glycan notation, numeric assignments, and clustering representations are as described in the legend to Figure 5 in the main text.

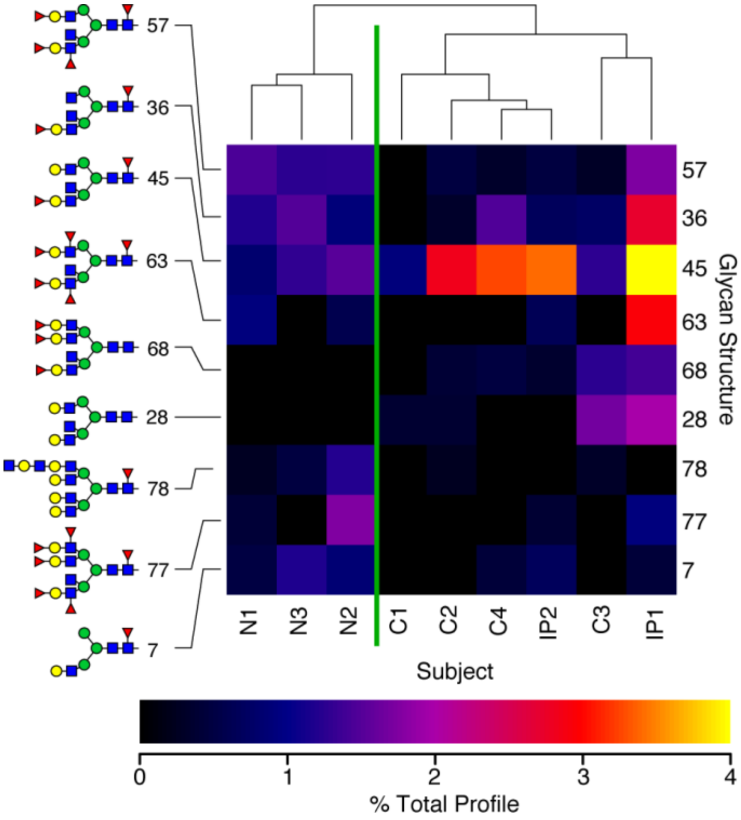


Figure 3.6. N-linked glycans that differentiate between normal and cancer/IPMN identified from whole glycan profiles.

A set of 6 dominant glycans (see Fig. 3.7) was removed from the whole profile and the prevalence of the remaining 79 glycans was recalculated for each sample. After recalculation, 9 glycans exhibited statistically significant changes comparing cancer (C1 - C4) and IPMN (IP1 – IP2) patients to normal (N1 – N3, Wilcoxon rank-sum $p \leq 0.05$). Hierarchical clustering of the prevalences of these 9 glycans demonstrates that their prevalences segregate normal ductal fluid glycan profiles from the glycan profiles of cancer or IPMN. Graphic representation of glycan structures are in accordance with the guidelines proposed by the Consortium for Functional Glycomics (CFG): blue square, N-acetylglucosamine (GlcNAc); green circle, mannose (Man); yellow circle, galactose (Gal); red triangle, fucose (Fuc); pink diamond, sialic acid as N-acetylneuraminic acid (NeuAc); light blue diamond, sialic acid as N-glycolylneuraminic acid (NeuGc). Glycan numbers are provided as arbitrary identifiers and refer consistently to the same structure throughout the manuscript and in the supplementary information (figures and tables). Brackets across the top of the cluster diagram provide a graphic presentation of the relatedness of the profile defined by each column. Thus, the total path length separating any two samples is directly proportional to the similarity of the glycan profile presented by those samples. For instance, N1 and N3 are more similar to each other than either is to N2 and all of the N samples are more similar to each other than any of the IP or C samples.

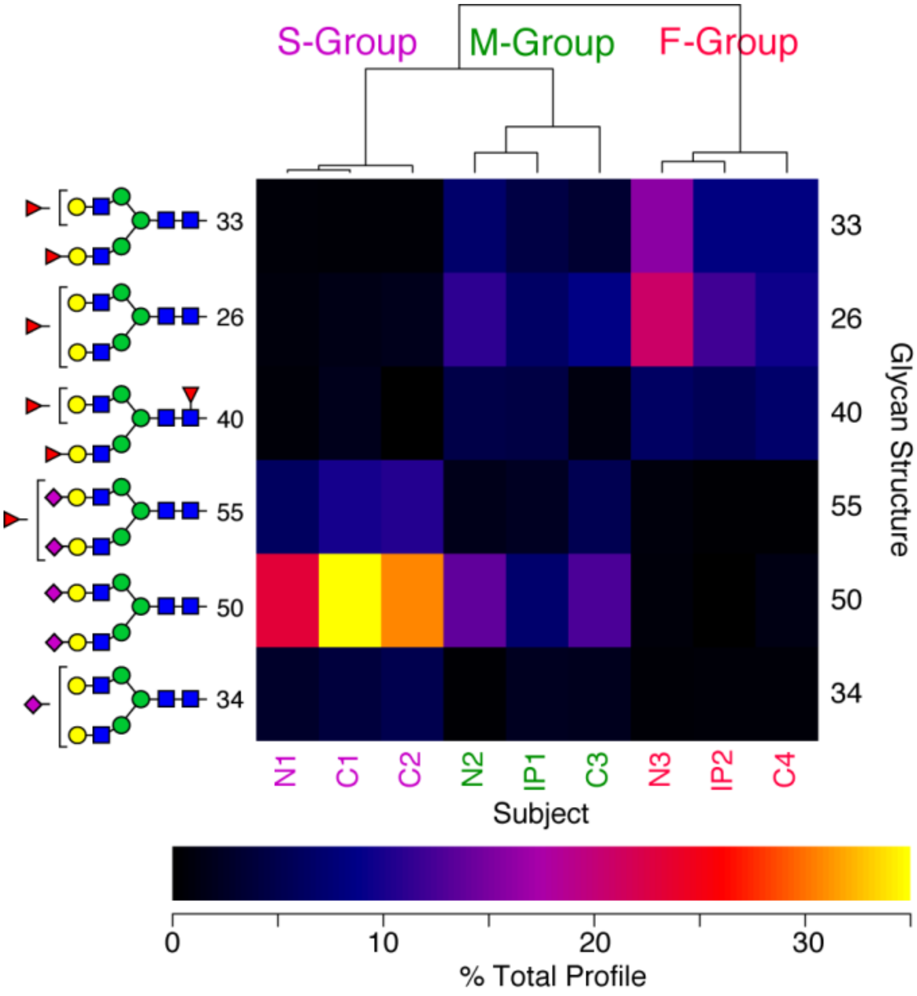


Figure 3.7. Dominant glycans define three distinct sample groups.
Analysis of 9 pancreatic ductal fluid samples identified 6 glycans that dominate the total glycan profiles of discrete sample subsets. These driver glycans defined three groups: S-group, dominated by sialylated glycans; F-group, dominated by fucosylated glycans; and M-group, presented a balance of the S- and F-group drivers. Hierarchical clustering robustly segregates S-, M-, and F-group samples. N1, N2, and N3: normal samples. C1, C2, C3, and C4: cancer samples. IP1, and IP2: IPMN samples. Glycan notation, numeric assignments, and clustering representations are as described in the legend to Figure 5.

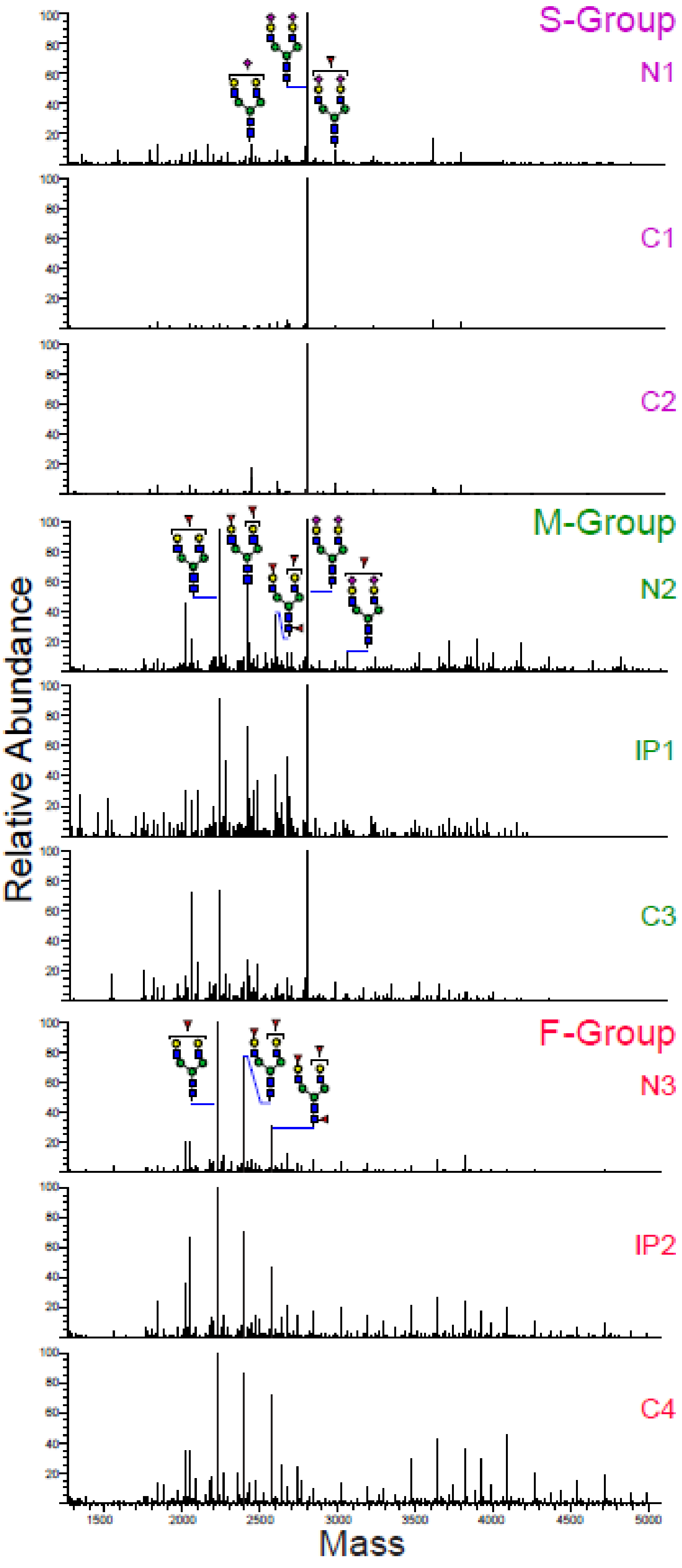


Figure 3.8. Deconvoluted mass spectra of N-glycans released from pancreatic ductal fluid. Full MS RAW data was collected by positive FT ion mode in an Orbitrap Disovery using nanospray direct infusion. The resulting data was de-isotoped and deconvoluted by the Xtract functionality of the Xcalibur data package (Thermo Fisher Scientific). Three patterns of dominant glycan profiles were observed, defining the S-Group, M-Group, and F-Group samples. N, normal; C, cancer; IP, IPMN. Peaks corresponding to the driver glycans for each class are annotated by structure.

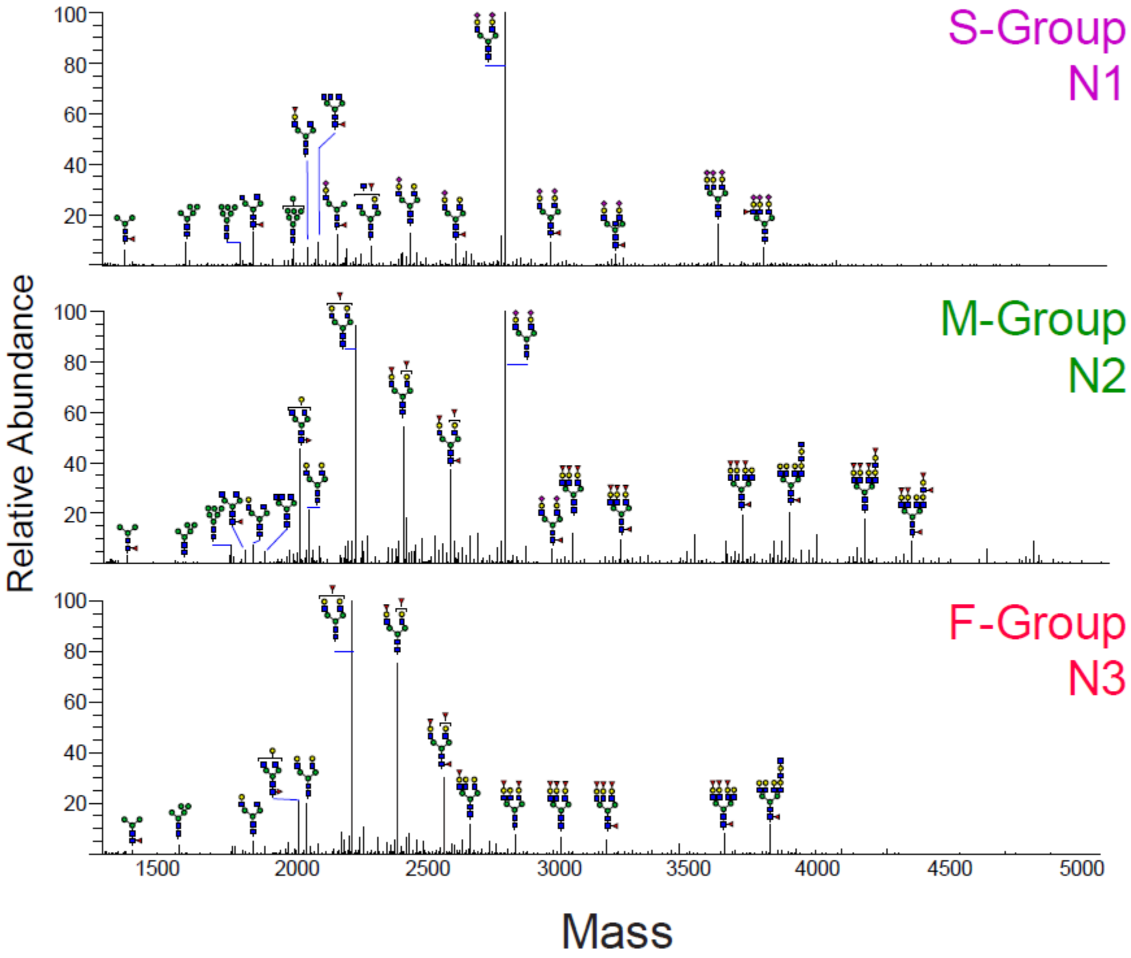


Figure 3.9. Additional annotation of mass spectra of N-glycans released from pancreatic ductal fluid harvested from normal patients representing the S-, M-, and F-Groups. To facilitate the comparison of total N-linked glycomic differences between the groups, full MS spectra from one representative individual of each group was de-isotoped, deconvoluted, and annotated in greater detail than is presented in Figure 3.8.

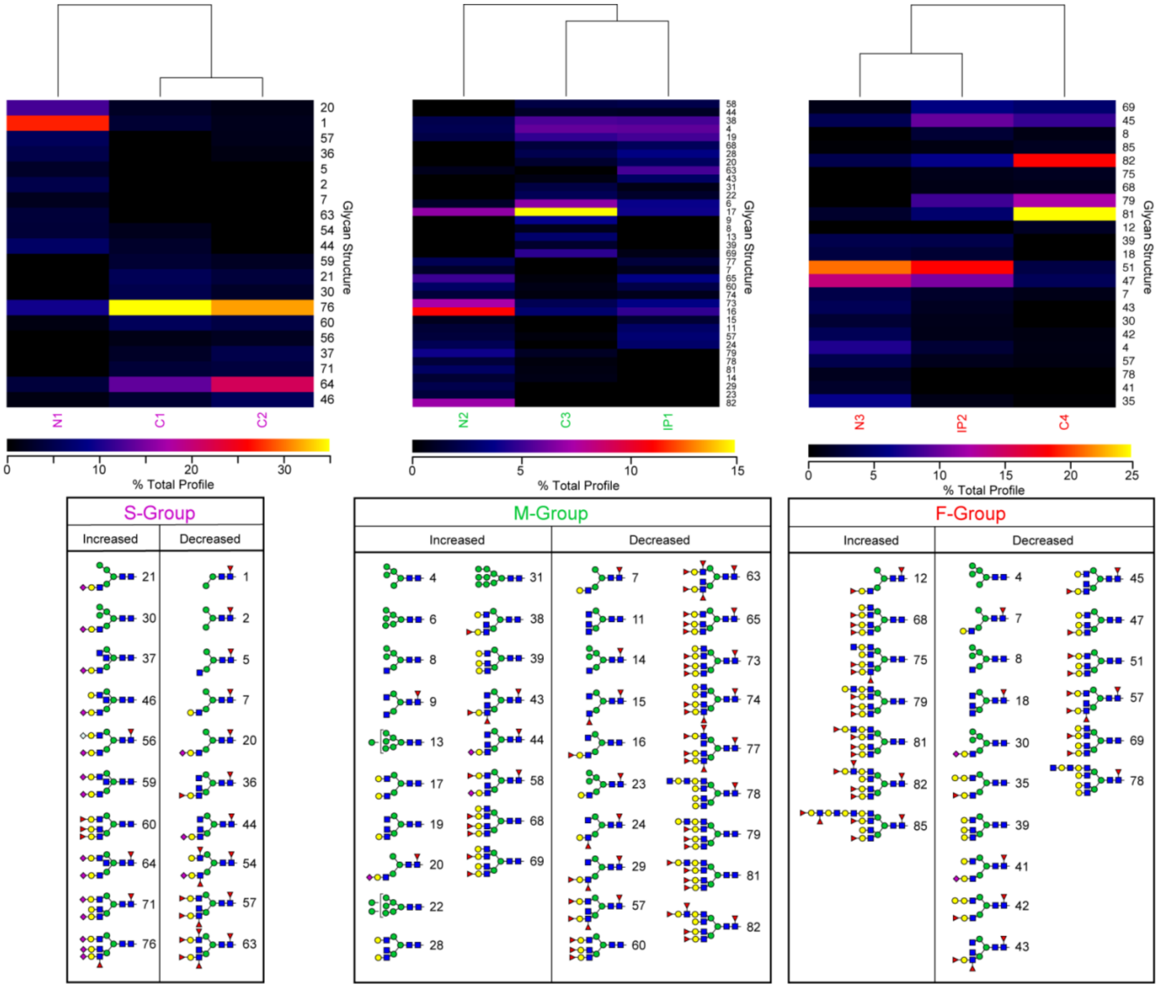


Figure 3.10. Glycan structures that distinguish cancer from normal within sample groups.

For each of the three groups defined by driver glycans (see Fig. 6), the 6 driver glycans were removed and the prevalence of the remaining glycans was recalculated. Subsequently, glycan prevalences were compared within each group. Glycans that did not distinguish between normal and cancer (≤ 2-fold increase or decrease) within a group were removed and the prevalences of the remaining glycans were recalculated. Hierarchical clustering of the residual glycan pool identified glycan subsets that were increased or decreased comparing normal to cancer. N1, N2, and N3: normal samples. C1, C2, C3, and C4: cancer samples. IP1, and IP2: IPMN samples. Glycan notation, numeric assignments, and clustering representations are as described in the legend to Figure 3.6.

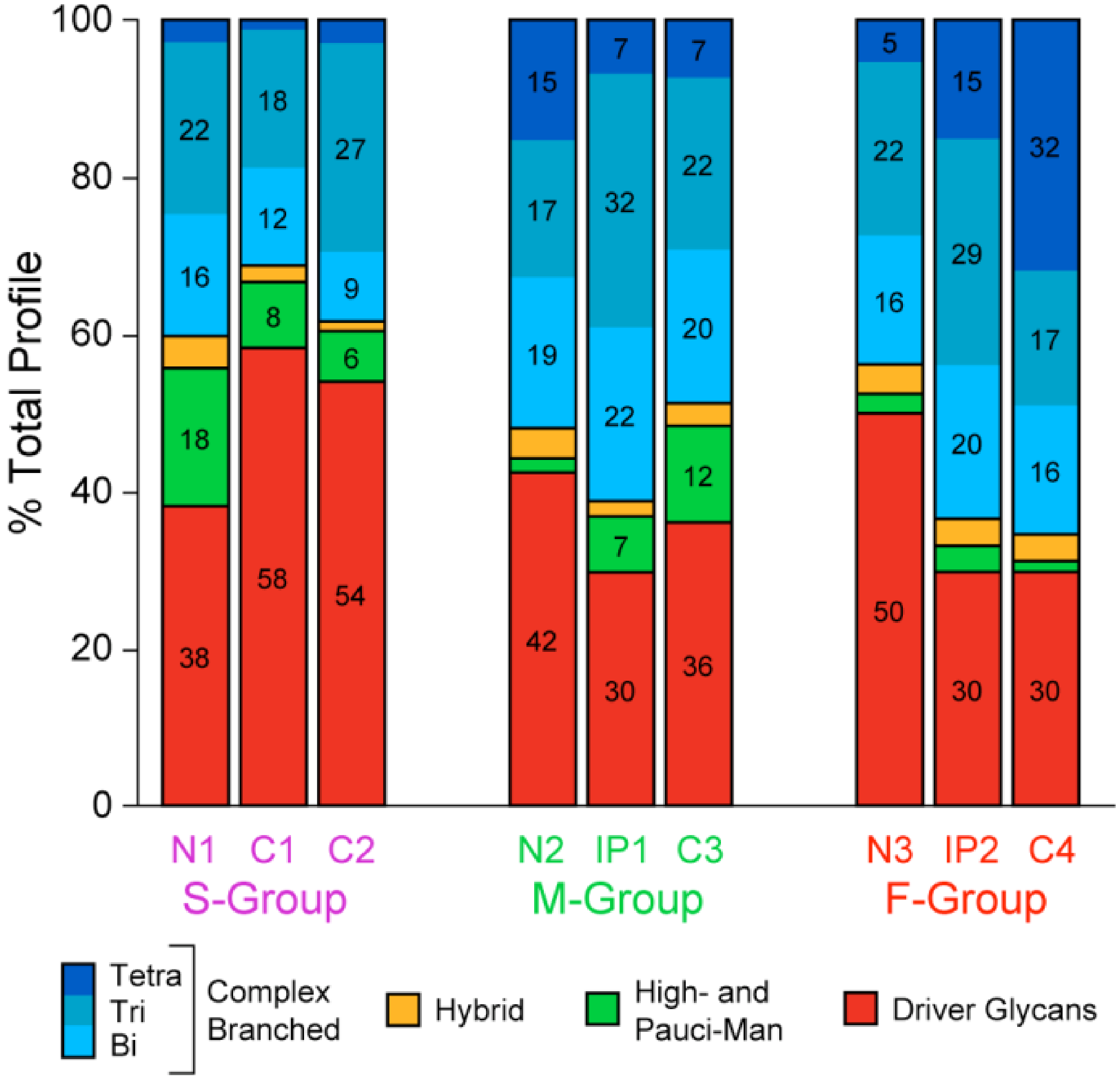


Figure 3.11. Summary of group-specific changes in glycan structural features that discriminate normal from cancer.

For S-group samples, cancer glycan profiles are characterized by increased driver glycan expression and decreased high-mannose glycans. For F-group samples, driver glycans are decreased in cancer samples and glycan complexity is increased. For M-group samples, driver glycan expression is not significantly altered, but high-mannose structures are increased and the most complex glycan structures are decreased in prevalence. Thus, each group exhibits unique changes in ductal fluid glycan expression when comparing normal to cancer. N1, N2, and N3: normal samples. C1, C2, C3, and C4: cancer samples. IP1, and IP2: IPMN samples.

Discussion

Glycomic and Proteomic Biomarker Strategy

The identification of proteomic markers for human disease holds promise for improving early diagnosis and for enhancing clinicians' ability to monitor treatment efficacy. Glycomic markers offer similar opportunities and several are currently in service as cancer diagnostics (CA125, CA19-9, core fucosylation of αFP). We have proposed that a marker or set of markers that reports changes in both protein and glycan composition could potentially yield higher sensitivity and specificity than a single protein or glycan marker. This is especially true in a disease that affects a small percentage of the population such as pancreatic cancer in the United States. With an incidence rate of approximately 1 in 10,000 Americans, a screen with a 1% false-positive rate and 1% false-negative rate of one million Americans would be expected to come back with 10,099 positive identifications. 99 of these would be expected to be true-positives (from an expected 100 in the population group) while 10,000 of them would be false-positives. This simple example illustrates the need for multiplexing and orthogonal analyses.

Predictive protein glycosylation changes might be found on proteins that are themselves biomarkers or might be detected as a change to the whole glycan profile of a biological sample. Here, we have concentrated on defining the scope of the proteomic and glycomic changes associated with pancreatic cancer, as detected in ductal fluid. Our quantitative proteomic results identified several proteins that are distinctively upregulated or downregulated in pancreatic cancer and/or IPMN samples, and some proteins that are only detected in cancer. Our analysis of N-linked glycans released from ductal fluid glycoproteins revealed unexpected diversity and interesting commonalities between subjects and also detected changes in glycan expression that correlate with pancreatic cancer. Future work will investigate whether the glycosylation changes that we have detected can be mapped to the putative glycoprotein markers that we have identified. The current work (using a training set of samples) establishes important parameters for validating candidate glycomic and proteomic biomarkers (using a confirmatory set of samples) and for interpreting and expanding this biomarker discovery effort.

Secreted Pancreatic Enzyme Proteomics

The ductal fluid proteins most predictive of cancer in our training set are primarily secreted proteins that possess degradative enzyme activities consistent with the exocrine function of the pancreas. A subset of lipases, glycosidases, and proteases exhibit changes ranging from a 6-fold increase to a 22-fold decrease in cancer. We detected decreased pancreatic amylase (AMYP) in cancer ductal fluid, consistent with previous studies in a rat model that demonstrated significant loss of amylase from pancreatic tumor cells by immunocytochemistry[28]. Likewise, decreased levels of elastase activity in duodenal aspirates have been reported following secretin-induced secretion in chronic pancreatitis, pancreatic cancer, and liver cirrhosis patients compared to normal controls[29]. Consistent with this result, and with another recent proteomic analysis of pancreatic ductal fluid by Gao, et al., we detected decreased elastase proteins (ELA2A, 3A, and 3B) in cancer[10]. However, increased levels of elastase 3b (formerly designated elastase 1) in pancreatic cancer tissue samples have been observed by several groups[30-35]. Considering the full range of proteomic analyses completed to date, the utility of elastase 3b as a biomarker for pancreatic cancer is in doubt, especially in light of its reported changes in pancreatitis as well as cancer. Pancreatic lipase-related protein 2 (LIPR2), which is the major colipase-dependent lipase in the pancreas[36], has been implicated in tumor cell killing through apoptotic and necrotic death induced by high levels of unsaturated fatty acids[37-44]. Consistent with the proteomic results presented here, pancreatic lipase immunoreactivity in serum was shown to decrease in pancreatic cancer[45]. However, the recent proteomic characterization by Gao, et al. reported increased LIPR1 in cancer and failed to detect LIPR2, while we detect no change in LIPR1 and decreased LIPR2[10].

The proteomic analysis by Gao, et al. also reported increased serine proteinase-2 (PRSS2, trypsin 2) in cancer, a protein that we failed to detect, but our analysis did identify PRSS1 (trypsin 1) as a significantly increased candidate biomarker for cancer[10]. Disagreements between the current study and the results of Gao, et al. likely reflect the very different techniques used (2D-gel electrophoresis followed by MALDI-TOF/MS versus LC-iontrap MS/MS). Another protein that we measured to be increased in pancreatic cancer is phospholipase A2 (PLA2), which has been previously associated with breast, lung and prostate cancers[46-56]. We were unable to detect PLA2 in normal ductal fluid, although it was present in ductal fluid of cancer, IPMN, and pancreatitis patients, thereby failing to discriminate based on presence between pancreatic disease types. Therefore, among the major secreted enzymes detected in pancreatic ductal fluid, we have identified increased AMYP and PRSS1, as well as decreased LIPR2 as candidate biomarkers worthy of further validation.

Non-enzyme Pancreatic Proteomics

Non-enzyme proteins, such as GP2-1, CCDC132, and REG family members also show significant changes in pancreatic ductal fluid. GP2-1 is a major glycoprotein of pancreatic acinar cell secretory granules and our analysis demonstrates that it is significantly increased in ductal fluid of IPMN and cancer patients compared to normal[57]. GP2-1 exists as a GPI-anchored form and as a truncated form that is secreted into the ductal fluid. GP2-1 function is incompletely characterized but the protein has significant similarity to uromodulin (Tamm-Horsfall protein), a kidney protein secreted into the urine and associated with renal innate immunity and ionic homeostasis[58]. Similarly, the coiled-coil domain-containing protein 132 (CCDC132) was only quantifiable or detectable in cancer and in one IPMN sample in our dataset; it was not detected in normal ductal fluid. Like GP2-1, the function of CCDC132 is currently unknown, but this cytoplasmic phosphoprotein possesses an N-terminal domain with homology to vacuolar sorting factors, suggesting a role in protein trafficking and association with secretory or transport vesicles[59]. The REG proteins are a group of structurally related proteins that stimulate proliferation and differentiation of liver, pancreatic, gastric, and intestinal cell populations[27]. Members of the REG protein family have been linked to gastric, liver, and pancreatic cancer and we detected increased REG1a, REG1b, and REG3a proteins in ductal fluid from cancer patients[60-63]. Interestingly, we not only detected changes in REG protein amounts, but also in REG protein glycoform distribution by Western blotting. In general, REG proteins exhibited greater heterogeneity in cancer than in normal ductal fluid. The differences that generate this heterogeneity are currently uncharacterized and may reflect distinct glycosylation profiles or differential proteolyitic processing of the apoproteins. Therefore, among the non-enzyme proteins that we detected in pancreatic ductal fluid, we have identified increased GP2-1, CCDC132, REG1a, REG1b, and

REG3a, as well as increased heterogeneity of REG proteins as candidate biomarkers requiring further validation.

N-linked Glycomics of Pancreatic Cancer

Changes in N-linked protein glycosylation have been described in many cancers including pancreatic cancer, in which altered glycosylation of serum proteins has been demonstrated[64-67]. The serum proteins previously reported to exhibit altered N-linked glycosylation in pancreatic cancer are acute phase proteins normally produced by hepatocytes, not pancreatic cells, in response to systemic inflammation. By monitoring the glycan profile of pancreatic ductal fluid, we have accessed the secretory products of normal and cancerous pancreatic cells in the compartment of closest proximity to their biosynthetic origin. Our purpose was to assess glycan profiles in a biological sample that would provide the greatest opportunity to detect relevant changes. Surprisingly, our N-linked glycomic analysis revealed 3 distinct glycosylation signatures within the 9 samples analyzed. Each signature was defined by a set of highly prevalent driver glycans.

These signatures were independent of cancer diagnosis, blood group status, and any other characteristic captured by our sample collection protocol, including age and sex. However, the structural features of the driver glycans that define these groups suggest that expression of the Secretor α1-2 Fucosyltransfease (Se FucT), which generates blood group H epitopes in epithelial cells, might account for some of the differences. If true, the S-group could be assigned as *Se-*/*Se-*, the M-group as *Se-*/*Se+*, and the F-group as *Se+*/*Se+*. These assignments make sense when considering the reciprocal gradation of fucosylation/sialylation across the groups; decreasing Se FucT activity increases the prevalence of unmodified terminal Gal residues that are substrates for sialyltransferases. But the changes we observed in the total glycan profiles of cancer samples within each group indicate a more complex scenario. In particular, it is difficult to propose how altered terminal fucosylation might affect branching, which we detect as increased in the S- and F-groups but decreased in the M-group. Likewise, increased poly-LacNAc in the F-group was countered by decreased poly-LacNAc in the M-group. Such divergent changes in cancer glycan expression cannot be linked simply to the level of active Se FucT. Rather, the divergent glycan profiles must reflect underlying changes in protein glycosylation that accompany cancer progression.

It remains to be determined whether all human samples can be clustered into these three groups or whether additional glycan signatures will be defined by other principal components as we increase the statistical power of our analysis. No analogous balkanization of glycan profiles has been described for serum glycomics, perhaps reflecting the more restricted cellular origin of pancreatic ductal fluid in comparison to systemically circulating blood. Regardless of its origin, the sorting of human subjects into discrete glycomic bins provides unique opportunities to pursue personalized glycan-based diagnostics. Our data indicates that proteomic and glycomic analysis of pancreatic ductal fluid must first assign samples to a driver glycan class before attempting to decipher the relevance of candidate glycoproteomic markers. On one hand, this classification requires that more samples must be characterized to achieve statistical power within each population. But, once sorted, glycomic and proteomic differences may provide markers capable of discriminating clinical diagnoses with greater specificity and sensitivity than is currently available.

Conclusions

Increases (AMYP, PRSS1) and decreases (LIPR2) of secreted pancreatic enzymes and increases of non-enzyme pancreatic proteins (GP2-1, CCDC132, REG1a, 1b, 3a) were detected in cancerous pancreatic ductal fluid in comparison to normal pancreatic ductal fluid in a small training set of samples. In addition, heterogeneity of the REG proteins was also found to increase in cancerous ductal fluid. A comprehensive analysis of the N-linked glycome of pancreatic ductal fluid identified unexpected clustering of patient samples into discrete subgroups that are enriched in sialylation or fucosylation, or are mixed with respect to both types of glycans independent of diagnosis. Within each group, changes in glycan prevalences are detected comparing normal to cancer albeit on a small number of samples. But, across groups, the glycan expression changes are different, even opposite in some cases. Therefore, interpretation of glycomic and glycoproteomic profiles must consider the heterogeneity of glycosylation across human populations before assessing the meaningfulness of changes in candidate biomarkers. The proteomic and glycomic features extracted from the training set of samples reported here establish important parameters for expanded validation and emphasize the need for large sample sets.

Table 3.1.
Number of identifiable and quantifiable proteins in pancreatic ductal fluid.

Identified Proteins				
Diagnosis	Proteins	Peptides	Spectral counts	Single-hit proteins
Cancer	213	1094	18206	77
IPMN[a]	149	831	22641	57
Pancreatitis	163	769	6645	61
Normal	136	775	11635	49
Combined	451	2082	59127	184
Quantified Proteins				
Diagnosis	Proteins	Peptides	Spectral counts	Unique proteins
Cancer	36	422	15092	7
IPMN	35	414	18632	8
Pancreatitis	19	215	3774	1
Normal	22	300	8674	0

[a] IPMN, Intraductal papillary mucinous neoplasm

Table 3.2.
List of proteins identified in pancreatic ductal fluid samples from different diagnoses.
All proteins identified with a protein FDR of <1% for each of the 4 diagnoses.
*table is too large for this thesis, please see original document at *Journal of Proteome Research* 2014 *13* (2), 395-407

Table 3.3.
List of peptides identified in pancreatic ductal fluid samples.
The dataset was filtered under a protein FDR of 1%. The presented parameters include the Xcorr scores, mass differences, charge states, and spectral counts for each peptide.
*table is too large for this thesis, please see original document at *Journal of Proteome Research* 2014 *13* (2), 395-407

Table 3.4.
Quantified pancreatic ductal fluid proteins differentially expressed in pancreatitis, IPMN, and cancer relative to normal.

Uniprot Accession	Abbreviation	Protein Name	Gene Name	Protein Length (AA)	Protein Weight (kDa)	PT/N[a] (log2 of ratio)	IP/N (log2 of ratio)	C/N (log2 of ratio)
P02787	TRFE	Serotransferrin	TF	698	76.982	3.49	NQ[b] in IP	2.54
P04118	COL	Colipase	CLPS	112	11.928		0.44	0.04
P04745	AMY1	Alpha-amylase 1	AMY1A	511	57.713	-0.60	2.30	2.95
P04746	AMYP	Pancreatic alpha-amylase	AMY2A	511	57.652	1.05	-0.24	-1.96
P05451	REG1A	Lithostathine-1-alpha	REG1A	166	18.701	0.23	0.52	1.13
P07477	TRY1	Trypsin-1	PRSS1	247	26.523	0.37	1.11	1.48
P08217	ELA2A	Elastase-2A	ELA2A	269	28.851	-0.33	-0.29	-1.65
P08861	ELA3B	Elastase-3B	ELA3B	270	29.256	-2.39	-1.17	-1.93
P09093	ELA3A	Elastase-3A	ELA3A	270	29.438	-0.23	-0.75	-0.67
P15085	CBPA1	Carboxypeptidase A1	CPA1	419	47.093	-0.39	-0.73	-0.70
P15086	CBPB1	Carboxypeptidase B	CPB1	417	47.320	-0.32	-0.80	0.08
P16233	LIPP	Pancreatic triacylglycerol lipase	PNLIP	465	51.106	0.57	0.88	0.18
P17538	CTRB1	Chymotrypsinogen B	CTRB1	263	27.834	0.54	0.86	0.78
P19835-1	CEL	Isoform Long of Bile salt-activated lipase	CEL	742	78.278	-0.01	-0.19	-0.72
P19961	AMY2B	Alpha-amylase 2B	AMY2B	511	57.655	NQ in PT	1.30	NQ in C
P48052	CBPA2	Carboxypeptidase A2	CPA2	417	46.781	NQ in PT	0.31	-0.54
P54315-1	LIPR1	Isoform 1 of Pancreatic lipase-related protein 1	PNLIPRP1	467	51.797	NQ in PT	-0.20	NQ in C
P54317	LIPR2	Pancreatic lipase-related protein 2	PNLIPRP2	469	51.895	NQ in PT	NQ in IP	-4.56
P55259-1	GP2-1	Isoform 1 of Pancreatic secretory granule membrane major glycoprotein GP2	GP2	537	59.424	NQ in PT	2.87	3.73
P55259-3	GP2-3	Isoform Alpha of Pancreatic secretory granule membrane major glycoprotein GP2	GP2	534	59.071	NQ in PT	1.37	-0.65
Q3SY19	PRSS1	PRSS1 protein	PRSS1	247	26.521	-0.20	0.50	1.98

[a] N: Normal; PT: Pancreatitis; IP: Intraductal papillary mucinous neoplasm; C: Cancer.
[b] NQ: Not quantifiable in the indicated diagnosis.

Table 3.5.
Quantified pancreatic ductal fluid proteins differentially expressed in IPMN and cancer relative to pacreatitis.

Uniprot Accession	Abbreviation	Protein Name	Gene Name	Protein Length (AA)	Protein Weight (kDa)	IP/PT[a] (log2 of ratio)	C/PT (log2 of ratio)
P68871	HBB	Hemoglobin subunit beta	HBB	147	15.970	-2.89	-1.00
A8K008	A8K008	cDNA FLJ78387	NA[b]	472	51.546	NQ[c] in IP	-2.84
Q5EFE6	Q5EFE6	Anti-RhD monoclonal T125 kappa light chain	NA	234	25.664	NQ in IP	-0.57

[a] PT: Pancreatitis; IP: Intraductal papillary mucinous neoplasm; C: Cancer.
[b] NA: None assigned.
[c] NQ: Not quantifiable in the indicated diagnosis.

Table 3.6.
Pancreatic ductal fluid proteins unique to each diagnosis.

			Quantifiable In		Identifiable In			
			C[a]	IP	C	IP	PT	N
Abbreviation	Protein Name	Gene Name	(# Patients)		(# Patients)			
CO3	Complement C3	C3	2	0	3	0	1	0
IGHG3	Ig gamma-3 chain C region	IGHG3	2	0	2	1	0	0
REG1B	Lithostathine-1-beta	REG1B	2	0	2	1	1	1
CC132	Isoform 1 of Coiled-coil domain-containing protein 132	CCDC132	2	0	2	1	0	0
A0A5E4	Putative uncharacterized protein	NA[b]	2	0	2	1	1	1
Q569I7	Putative uncharacterized protein	NA	2	0	2	1	1	0
Q6ZP64	CDNA FLJ26451 fis, clone KDN03041	NA	2	0	2	0	0	0
PA21B	Phospholipase A2	PLA2G1B	2	3	2	1	2	0
ELA2B	Elastase-2B	ELA2B	2	3	3	1	1	0
TRY3	Isoform A of Trypsin-3	PRSS3	2	3	3	1	1	0
HBA	Hemoglobin subunit alpha	HBA1	2	3	3	1	0	0
REG3A	Regenerating islet-derived protein 3 alpha	REG3A	2	2	2	0	0	0
CTRC	Chymotrypsin-C	CTRC	2	3	3	1	1	0

[a] N: Normal; PT: Pancreatitis; IP: Intraductal papillary mucinous neoplasm; C: Cancer.
[b] NA: None assigned.

Table 3.7.
List of proteins quantified in pancreatic ductal fluid samples from different diagnoses. Proteins with a protein FDR of <1% were further filtered at peptide FDR of <10%. Only proteins identified by more than one peptide that were present in more than one sample with the same diagnosis were retained. *table is too large for this thesis, please see original document at *Journal of Proteome Research* 2014 *13* (2), 395-407

Table 3.8.
Patient information for samples used in this study.
The approved protocol for collecting patient information for this study captured diagnosis, age, and sex for each patient. Due to limited sample amount, not all samples could be analyzed by all assays. The analyses performed on each sample are indicated.

Diagnosis	Patient identifier	Age	Sex	Analysis*
Normal	N1	72	male	Gly
	N2	42	female	Gly, Pro
	N3	63	female	Gly, Pro
	N4	64	female	WB
	N5	47	male	WB
	N6	34	female	WB
	N7	33	female	WB
	N8	31	female	Pro, WB
Pancreatitis	PT1	56	male	Pro
	PT2	41	female	Pro
	PT3	47	female	Pro
IPMN	IP1	82	male	Gly, Pro
	IP2	72	male	Gly, Pro
	IP3	77	female	Pro
Cancer	C1	79	male	Gly
	C2	61	female	Gly
	C3	60	female	Gly
	C4	54	male	Gly
	C5	60	male	WB
	C6	72	female	Pro, WB
	C7	66	male	Pro, WB
	C8	71	male	Pro, WB
	C9	66	female	Pro, WB

*Gly, Glycomics; Pro, Proteomics; WB, western blot

Table 3.9. N-linked glycans identified in pancreatic ductal fluid samples.
The monosaccharide composition and glycan prevalences are presented for each normal, cancer, and IPMN sample analyzed. Prevalences are presented as "% Total Profile," calculated by dividing the signal intensity at the indicated mass by the total intensity for all detected glycans.

		Monosaccharide Composition					% Total Profile								
Glycan #	Mol Weight	Fuc	NeuAc	Hex	HexNAc	NeuGc	N1	N2	N3	C1	C2	C3	C4	IP1	IP2
1	1141.6	1	0	2	2	0	5.15	0.00	0.27	0.36	0.29	0.36	0.00	0.00	0.00
2	1345.7	1	0	3	2	0	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	1560.8	2	0	2	3	0	0.00	0.00	0.00	0.00	0.15	0.00	0.00	1.56	0.00
4	1579.8	0	0	5	2	0	4.75	1.21	1.58	3.51	1.82	3.58	0.33	2.82	0.74
5	1590.8	1	0	3	3	0	0.50	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00
6	1783.9	0	0	6	2	0	3.77	0.61	0.29	3.11	3.06	4.09	0.34	1.93	1.84
7	1794.9	1	0	4	3	0	0.38	0.61	0.73	0.00	0.00	0.00	0.38	0.37	0.58
8	1824.9	0	0	5	3	0	0.25	0.00	0.00	0.00	0.00	0.45	0.33	0.00	0.74
9	1835.9	1	0	3	4	0	4.66	0.00	0.00	4.33	3.59	2.32	0.00	0.00	0.00
10	1865.9	0	0	4	4	0	0.00	0.00	1.14	0.16	0.00	0.00	1.84	0.36	3.22
11	1907.0	0	0	3	5	0	0.20	0.79	0.71	0.00	0.54	0.00	1.19	1.36	0.81
12	1969.0	2	0	4	3	0	0.65	0.00	0.00	0.50	0.00	0.00	0.58	0.35	0.00
13	1988.0	0	0	7	2	0	0.71	0.00	0.00	0.43	0.71	1.77	0.00	0.00	0.00
14	1999.9	1	0	5	3	0	0.00	0.77	0.89	0.19	0.00	0.37	0.78	0.00	0.86
15	2010.0	2	0	3	4	0	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.71	0.00
16	2040.0	1	0	4	4	0	1.62	6.04	4.60	1.87	0.88	1.80	3.44	2.33	4.28
17	2070.0	0	0	5	4	0	0.45	3.74	4.66	0.27	0.23	8.99	5.73	1.97	8.14
18	2081.0	1	0	3	5	0	1.92	0.66	0.63	0.79	0.72	1.18	0.00	1.09	0.71
19	2111.1	0	0	4	5	0	0.41	1.04	0.91	0.00	0.43	2.88	1.56	2.53	0.96
20	2156.1	1	1	4	3	0	2.33	0.00	0.00	0.26	0.23	0.79	0.00	1.26	0.00
21	2186.1	0	1	5	3	0	0.00	0.43	0.00	0.60	0.60	0.43	0.00	0.00	0.00
22	2192.1	0	0	8	2	0	1.38	0.00	0.35	0.45	0.43	1.42	0.48	0.58	0.49
23	2203.1	1	0	6	3	0	0.00	0.90	1.60	0.00	0.00	0.00	1.34	0.00	0.87
24	2214.1	2	0	4	4	0	0.31	1.25	0.94	0.00	0.00	0.00	1.55	1.41	1.35
25	2227.1	0	1	4	4	0	0.56	0.53	0.42	0.56	0.87	0.45	0.33	0.60	0.00
26	2244.1	1	0	5	4	0	1.03	12.33	23.19	1.55	2.00	9.43	10.64	7.24	13.49
27	2285.2	1	0	4	5	0	1.93	1.28	2.29	1.20	0.69	2.61	1.83	3.67	1.76

28	2316.2	0	0	5	5	0	0.00	0.00	0.00	0.19	0.21	1.33	0.00	1.74	0.00
29	2388.2	3	0	4	4	0	0.00	1.15	0.45	0.00	0.00	0.00	0.46	0.00	0.00
30	2390.2	0	1	6	3	0	0.00	1.15	0.51	0.53	0.42	0.59	0.00	0.00	0.38
31	2396.2	0	0	9	2	0	0.90	0.00	0.00	0.51	0.00	1.10	0.29	0.31	0.30
32	2401.2	1	1	4	4	0	1.29	0.97	0.36	0.47	0.00	0.64	0.38	0.57	0.00
33	2418.2	2	0	5	4	0	0.65	7.63	17.59	0.59	0.47	3.56	9.21	4.99	9.15
34	2431.2	0	1	5	4	0	3.16	0.44	0.50	4.50	5.71	2.15	0.73	2.40	0.69
35	2448.2	1	0	6	4	0	0.04	0.83	1.39	0.11	0.00	0.77	0.09	0.96	0.31
36	2459.2	2	0	4	5	0	0.90	0.64	0.92	0.00	0.18	0.59	1.30	2.38	0.59
37	2473.2	0	1	4	5	0	0.00	0.00	0.00	0.28	0.76	0.00	0.00	0.00	0.00
38	2489.2	1	0	5	5	0	0.70	1.15	1.34	0.48	0.36	3.25	1.46	2.58	1.44
39	2519.3	0	0	6	5	0	0.00	0.00	0.76	0.21	0.18	0.76	0.00	0.00	1.04
40	2592.3	3	0	5	4	0	0.63	5.28	7.08	1.98	0.00	1.13	7.64	4.93	6.21
41	2605.3	1	1	5	4	0	2.10	1.74	0.44	2.89	1.43	1.40	0.00	4.05	0.00
42	2622.3	2	0	6	4	0	0.19	0.44	0.85	0.00	0.00	0.25	0.28	0.00	0.35
43	2633.0	3	0	4	5	0	0.36	0.00	0.84	1.50	0.19	0.31	0.00	1.32	0.44
44	2646.3	1	1	4	5	0	1.21	0.00	0.00	0.31	0.00	0.68	0.00	0.49	0.00
45	2663.3	2	0	5	5	0	0.62	1.08	0.80	0.47	1.62	1.01	2.87	3.48	2.98
46	2676.3	0	1	5	5	0	0.21	0.00	0.00	0.34	0.92	0.00	0.00	0.00	0.00
47	2693.3	1	0	6	5	0	0.00	1.90	3.06	0.00	0.00	1.38	1.42	0.83	3.22
48	2767.0	4	0	5	4	0	0.86	0.95	1.06	0.14	0.67	0.89	2.55	8.44	2.07
49	2778.4	2	1	5	4	0	2.77	1.33	0.21	1.34	0.92	1.86	0.00	0.67	0.00
50	2792.4	0	2	5	4	0	25.75	15.23	0.78	38.73	34.24	14.04	1.38	7.87	0.00
51	2868.4	2	0	6	5	0	0.30	1.18	4.59	0.08	0.09	1.27	1.10	1.33	5.57
52	2891.4	1	1	4	6	0	1.27	0.00	0.00	1.21	1.21	0.37	0.00	0.00	0.00
53	2908.0	2	0	5	6	0	0.19	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.43
54	2952.5	3	1	5	4	0	0.70	0.00	0.00	0.20	0.00	0.29	0.00	0.26	0.00
55	2966.5	1	2	5	4	0	6.92	1.95	1.01	10.93	11.78	5.87	0.44	2.51	0.42
56	2997.5	1	1	5	4	1	0.00	0.00	0.00	0.15	0.33	0.00	0.00	0.00	0.00
57	3011.5	4	0	5	5	0	1.10	0.91	0.78	0.00	0.27	0.23	0.27	1.51	0.41
58	3024.5	2	1	5	5	0	1.20	0.00	0.00	0.15	0.53	1.14	0.00	1.02	0.00
59	3036.9	0	2	5	5	0	0.00	0.00	0.00	0.37	0.35	0.00	0.00	0.00	1.90
60	3042.6	3	0	6	5	0	0.22	1.49	1.61	0.62	0.69	0.36	1.70	0.95	3.10
61	3082.6	3	0	5	6	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00

62	3142.6	1	0	7	6	0	0.00	0.00	0.00	0.00	0.33	0.00	0.28	0.35	0.26
63	3186.6	5	0	5	5	0	0.71	0.42	0.00	0.00	0.00	0.00	0.00	2.56	0.55
64	3211.6	1	2	5	5	0	0.74	0.19	0.06	1.48	3.41	0.25	0.00	0.49	0.00
65	3215.6	4	0	6	5	0	0.00	2.72	2.86	0.00	0.00	0.19	2.62	1.76	3.56
66	3228.6	2	1	6	5	0	0.58	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00
67	3241.6	0	2	6	5	0	1.04	0.00	0.00	0.92	1.45	0.29	0.00	0.29	0.00
68	3285.6	3	0	6	6	0	0.00	0.00	0.00	0.00	0.23	1.00	0.41	1.22	0.31
69	3316.7	2	0	7	6	0	0.00	0.00	0.23	0.00	0.00	2.81	1.74	0.29	1.77
70	3402.7	3	1	6	5	0	0.18	0.00	0.00	0.18	0.16	0.00	0.00	0.00	0.00
71	3416.7	1	2	6	5	0	0.00	0.00	0.00	0.39	0.64	0.00	0.00	0.00	0.00
72	3602.8	0	3	6	5	0	5.50	1.88	0.00	3.54	7.27	1.70	0.00	0.55	0.00
73	3663.8	4	0	7	6	0	0.06	4.04	2.92	0.00	0.00	1.39	7.70	1.78	4.69
74	3693.8	3	0	8	6	0	0.24	0.38	0.21	0.00	0.00	0.00	0.00	0.00	0.00
75	3735.9	3	0	7	7	0	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.39	0.37
76	3776.6	1	3	6	5	0	1.82	0.35	0.00	3.79	4.99	0.53	0.00	0.68	0.00
77	3809.0	6	0	6	6	0	0.32	1.25	0.00	0.00	0.00	0.00	0.00	0.81	0.33
78	3836.9	1	0	8	8	0	0.19	0.86	0.30	0.00	0.13	0.24	0.00	0.00	0.00
79	3939.8	3	0	8	7	0	0.00	2.17	0.00	0.00	0.22	0.63	4.33	0.00	2.53
80	4084.0	5	0	7	7	0	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.73	0.00
81	4114.1	4	0	8	7	0	0.00	1.46	0.44	0.00	0.00	0.30	8.66	0.00	1.51
82	4288.1	5	0	8	7	0	0.00	3.84	0.76	0.00	0.00	0.00	6.48	0.00	1.90
83	4413.2	0	4	7	6	0	0.47	0.00	0.00	0.00	0.37	0.00	0.00	0.00	0.00
84	4634.3	3	0	9	9	0	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.17
85	5186.6	5	0	10	9	0	0.00	0.32	0.08	0.00	0.00	0.00	0.51	0.00	0.24

Table 3.10. Recalculated prevalences for each sample group.
After removal of driver glycans and glycans that do not exhibit differences between normal and cancer, prevalences were recalculated for the residual glycan population for each sample group. The resulting prevalences are expressed as "% Subset of Profile."

S-Group					**M-Group**			
	% Subset of Profile					% Subset of Profile		
Glycan#	N1	S1	S2		Glycan#	N2	C3	IP1
1	30.50	3.76	2.04		4	3.05	8.05	7.84
2	5.39	0.00	0.00		6	1.52	9.21	5.35
5	2.96	0.00	0.00		7	1.54	0.00	1.04
7	2.23	0.00	0.00		8	0.00	1.02	0.00
20	13.83	2.71	1.62		9	0.00	5.21	0.00
21	0.00	6.23	4.28		11	1.99	0.00	3.77
30	0.00	5.51	2.98		13	0.00	3.98	0.00
36	5.35	0.00	1.28		14	1.94	0.84	0.00
37	0.00	2.86	5.43		15	1.17	0.00	1.97
44	7.18	3.16	0.00		16	15.20	4.06	6.48
46	1.27	3.55	6.55		17	9.40	20.23	5.47
54	4.12	2.03	0.00		19	2.62	6.48	7.04
56	0.00	1.57	2.33		20	0.00	1.79	3.50
57	6.49	0.00	1.90		22	0.00	3.19	1.60
59	0.00	3.81	2.52		23	2.27	0.00	0.00
60	1.31	6.43	4.90		24	3.15	0.00	3.91
63	4.20	0.00	0.00		28	0.00	2.99	4.84
64	4.36	15.29	24.23		29	2.90	0.00	0.00
71	0.00	4.02	4.55		31	0.00	2.48	0.85
76	10.81	39.07	35.38		38	2.89	7.31	7.17
					39	0.00	1.70	0.00
F-Group					43	0.00	0.70	3.67
	% Subset of Profile				44	0.00	1.52	1.35
Glycan#	N3	C4	IP2		57	2.29	0.51	4.21
4	8.45	1.09	2.83		58	0.00	2.56	2.82
7	3.89	1.26	2.24		60	3.74	0.81	2.64
8	0.00	1.10	2.84		63	1.05	0.00	7.13
12	0.00	1.90	0.00		65	6.84	0.44	4.90
18	3.38	0.00	2.72		68	0.00	2.25	3.39
30	2.70	0.00	1.44		69	0.00	6.33	0.80
35	7.41	0.29	1.18		73	10.17	3.12	4.94
39	4.03	0.00	3.98		74	2.14	0.61	1.07
41	2.32	0.00	0.00		77	3.15	0.00	2.25
42	4.53	0.91	1.34		78	2.15	0.53	0.00
43	4.49	0.00	1.71		79	5.46	1.42	0.00
45	4.25	9.46	11.41		81	3.67	0.67	0.00
47	16.31	4.69	12.34		82	9.67	0.00	0.00
51	24.48	3.63	21.36					
57	4.13	0.91	1.56					
68	0.00	1.36	1.17					
69	1.22	5.74	6.79					
75	0.00	1.69	1.41					
78	1.57	0.00	0.00					
79	0.00	14.31	9.72					
81	2.37	28.59	5.79					
82	4.06	21.38	7.28					
85	0.41	1.69	0.90					

References

1. American Cancer Society Cancer Facts & Figures 2010. *Atlanta: American Cancer Society* 2010, 2010, 1-66.

2. Bernhard, J.; Dietrich, D.; Glimelius, B.; Hess, V.; Bodoky, G.; Scheithauer, W.; Herrmann, R., Estimating prognosis and palliation based on tumour marker CA 19-9 and quality of life indicators in patients with advanced pancreatic cancer receiving chemotherapy. *Br J Cancer* 2010, 103, (9), 1318-24.

3. Magnani, J. L.; Steplewski, Z.; Koprowski, H.; Ginsburg, V., Identification of the gastrointestinal and pancreatic cancer-associated antigen detected by monoclonal antibody 19-9 in the sera of patients as a mucin. *Cancer Res* 1983, 43, (11), 5489-92.

4. Xue, A.; Scarlett, C. J.; Chung, L.; Butturini, G.; Scarpa, A.; Gandy, R.; Wilson, S. R.; Baxter, R. C.; Smith, R. C., Discovery of serum biomarkers for pancreatic adenocarcinoma using proteomic analysis. *Br J Cancer* 2010, 103, (3), 391-400.

5. Kopp, J. L.; von Figura, G.; Mayes, E.; Liu, F. F.; Dubois, C. L.; Morris, J. P. t.; Pan, F. C.; Akiyama, H.; Wright, C. V.; Jensen, K.; Hebrok, M.; Sander, M., Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. *Cancer Cell* 2012, 22, (6), 737-50.

6. Chen, R.; Brentnall, T. A.; Pan, S.; Cooke, K.; Moyes, K. W.; Lane, Z.; Crispin, D. A.; Goodlett, D. R.; Aebersold, R.; Bronner, M. P., Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer. *Mol Cell Proteomics* 2007, 6, (8), 1331-42.

7. Chen, R.; Pan, S.; Cooke, K.; Moyes, K. W.; Bronner, M. P.; Goodlett, D. R.; Aebersold, R.; Brentnall, T. A., Comparison of pancreas juice proteins from cancer versus pancreatitis using quantitative proteomic analysis. *Pancreas* 2007, 34, (1), 70-9.

8. Chen, R.; Pan, S.; Yi, E. C.; Donohoe, S.; Bronner, M. P.; Potter, J. D.; Goodlett, D. R.; Aebersold, R.; Brentnall, T. A., Quantitative proteomic profiling of pancreatic cancer juice. *Proteomics* 2006, 6, (13), 3871-9.

9. Doyle, C. J.; Yancey, K.; Pitt, H. A.; Wang, M.; Bemis, K.; Yip-Schneider, M. T.; Sherman, S. T.; Lillemoe, K. D.; Goggins, M. D.; Schmidt, C. M., The Proteome of Normal Pancreatic Juice. *Pancreas* 2011.

10. Gao, J.; Zhu, F.; Lv, S.; Li, Z.; Ling, Z.; Gong, Y.; Jie, C.; Ma, L., Identification of pancreatic juice proteins as biomarkers of pancreatic cancer. *Oncol Rep* 2010, 23, (6), 1683-92.

11. Gronborg, M.; Bunkenborg, J.; Kristiansen, T. Z.; Jensen, O. N.; Yeo, C. J.; Hruban, R. H.; Maitra, A.; Goggins, M. G.; Pandey, A., Comprehensive proteomic analysis of human pancreatic juice. *J Proteome Res* 2004, 3, (5), 1042-55.

12. Rosty, C.; Goggins, M., Identification of differentially expressed proteins in pancreatic cancer using a global proteomic approach. *Methods Mol Med* 2005, 103, 189-97.

13. Comunale, M. A.; Rodemich-Betesh, L.; Hafner, J.; Wang, M.; Norton, P.; Di Bisceglie, A. M.; Block, T.; Mehta, A., Linkage specific fucosylation of alpha-1-antitrypsin in liver cirrhosis and cancer patients: implications for a biomarker of hepatocellular carcinoma. *PLoS One* 2010, 5, (8), e12419.

14. Marrero, J. A.; Romano, P. R.; Nikolaeva, O.; Steel, L.; Mehta, A.; Fimmel, C. J.; Comunale, M. A.; D'Amelio, A.; Lok, A. S.; Block, T. M., GP73, a resident Golgi glycoprotein, is a novel serum marker for hepatocellular carcinoma. *J Hepatol* 2005, 43, (6), 1007-12.

15. Nakagawa, T.; Takeishi, S.; Kameyama, A.; Yagi, H.; Yoshioka, T.; Moriwaki, K.; Masuda, T.; Matsumoto, H.; Kato, K.; Narimatsu, H.; Taniguchi, N.; Miyoshi, E., Glycomic analyses of glycoproteins in bile and serum during rat hepatocarcinogenesis. *J Proteome Res* 2010, 9, (10), 4888-96.

16. Sato, Y.; Nakata, K.; Kato, Y.; Shima, M.; Ishii, N.; Koji, T.; Taketa, K.; Endo, Y.; Nagataki, S., Early recognition of hepatocellular carcinoma based on altered profiles of alpha-fetoprotein. *N Engl J Med* 1993, 328, (25), 1802-6.

17. Wandall, H. H.; Blixt, O.; Tarp, M. A.; Pedersen, J. W.; Bennett, E. P.; Mandel, U.; Ragupathi, G.; Livingston, P. O.; Hollingsworth, M. A.; Taylor-Papadimitriou, J.; Burchell, J.; Clausen, H., Cancer biomarkers defined by autoantibody signatures to aberrant O-glycopeptide epitopes. *Cancer Res* 2010, 70, (4), 1306-13.

18. Aoki, K.; Perlman, M.; Lim, J.; Cantu, R.; Wells, L.; Tiemeyer, M., Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo. *J Biol Chem* 2007, 282, (12), 9127-9142.

19. Harris, M. A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G. M.; Blake, J. A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J. T.; Hill, D. P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J. M.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R. S.; Sethuraman, A.; Theesfeld, C. L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S. Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E. M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.;

Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R., The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, 32, (Database issue), D258-61.

20. Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 2006, 5, (9), 2339-47.

21. Zybailov, B. L.; Florens, L.; Washburn, M. P., Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol Biosyst* 2007, 3, (5), 354-60.

22. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* 2010, 82, (6), 2272-81.

23. Ciucanu, I.; Kerek, F., A simple and rapid method for the permethylation of carbohydrates. *Carbohydrate Research* 1984, 131, 209-217.

24. Apte, A.; Meitei, N. S., Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. *Methods Mol Biol* 2010, 600, 269-81.

25. Ferguson, M. W.; Wycoff, K. L.; Ayers, A. R., Use of cluster analysis with monoclonal antibodies for taxonomic differentiation of phytopathogenic fungi and for screening and clustering antibodies. *Curr Microbiol* 1988, 17, 127-132.

26. Pattathil, S.; Avci, U.; Baldwin, D.; Swennes, A. G.; McGill, J. A.; Popper, Z.; Bootten, T.; Albert, A.; Davis, R. H.; Chennareddy, C.; Dong, R.; O'Shea, B.; Rossi, R.; Leoff, C.; Freshour, G.; Narra, R.; O'Neil, M.; York, W. S.; Hahn, M. G., A comprehensive toolkit of plant cell wall glycan-directed monoclonal antibodies. *Plant Physiol* 2010, 153, (2), 514-25.

27. De Reggi, M.; Capon, C.; Gharib, B.; Wieruszeski, J. M.; Michel, R.; Fournet, B., The glycan moiety of human pancreatic lithostathine. Structure characterization and possible pathophysiological implications. *Eur J Biochem* 1995, 230, (2), 503-10.

28. Hilmy, A. M.; Kandeel, K. M.; Selim, N. M., Pancreatic amylase as a tumour marker for pancreatic cancer. *Arch Geschwulstforsch* 1984, 54, (6), 475-82.

29. Mizuno, R.; Hayakawa, T.; Noda, A., Elastase secretion in pancreatic disease. *Am J Gastroenterol* 1985, 80, (2), 113-7.

30. Achilles, K.; Bednarski, P. J., Quantification of elastase-like activity in 13 human cancer cell lines and in an immortalized human epithelial cell line by RP-HPLC. *Biol Chem* 2003, 384, (5), 817-24.

31. Iwase, K.; Miyata, M.; Yamaguchi, T.; Kawaguchi, T.; Tanaka, Y.; Matsuda, H., Pancreatic ductal cell carcinoma producing pancreatic elastase 1. *J Surg Oncol* 1993, 54, (3), 199-202.

32. Ventrucci, M.; Pezzilli, R.; Gullo, L.; Plate, L.; Sprovieri, G.; Barbara, L., Role of serum pancreatic enzyme assays in diagnosis of pancreatic disease. *Dig Dis Sci* 1989, 34, (1), 39-45.

33. Wendorf, P.; Geyer, R.; Sziegoleit, A.; Linder, D., Localization and characterization of the glycosylation site of human pancreatic elastase 1. *FEBS Lett* 1989, 249, (2), 275-8.

34. Wendorf, P.; Linder, D.; Sziegoleit, A.; Geyer, R., Carbohydrate structure of human pancreatic elastase 1. *Biochem J* 1991, 278 (Pt 2), 505-14.

35. Wu, D.; Qian, J. M.; Deng, R. X.; Jiang, W. J.; Chen, Y. J.; Liu, X. H.; Lu, X. H., Evaluating the role of serum elastase 1 in the diagnosis of pancreatic cancer. *Chin J Dig Dis* 2006, 7, (2), 117-20.

36. D'Agostino, D.; Lowe, M. E., Pancreatic lipase-related protein 2 is the major colipase-dependent pancreatic lipase in suckling mice. *J Nutr* 2004, 134, (1), 132-4.

37. Cury-Boaventura, M. F.; Curi, R., Regulation of reactive oxygen species (ROS) production by C18 fatty acids in Jurkat and Raji cells. *Clin Sci (Lond)* 2005, 108, (3), 245-53.

38. Cury-Boaventura, M. F.; Gorjao, R.; de Lima, T. M.; Newsholme, P.; Curi, R., Comparative toxicity of oleic and linoleic acid on human lymphocytes. *Life Sci* 2006, 78, (13), 1448-56.

39. Cury-Boaventura, M. F.; Pompeia, C.; Curi, R., Comparative toxicity of oleic acid and linoleic acid on Jurkat cells. *Clin Nutr* 2004, 23, (4), 721-32.

40. Cury-Boaventura, M. F.; Pompeia, C.; Curi, R., Comparative toxicity of oleic acid and linoleic acid on Raji cells. *Nutrition* 2005, 21, (3), 395-405.

41. Finstad, H. S.; Heimli, H.; Kolset, S. O.; Drevon, C. A., Proliferation and types of killing of leukemia cell lines by very long chain polyunsaturated fatty acids. *Lipids* 1999, 34 Suppl, S107.

42. Finstad, H. S.; Myhrstad, M. C.; Heimli, H.; Lomo, J.; Blomhoff, H. K.; Kolset, S. O.; Drevon, C. A., Multiplication and death-type of leukemia cell lines exposed to very long-chain polyunsaturated fatty acids. *Leukemia* 1998, 12, (6), 921-9.

43. Heimli, H.; Finstad, H. S.; Drevon, C. A., Necrosis and apoptosis in lymphoma cell lines exposed to eicosapentaenoic acid and antioxidants. *Lipids* 2001, 36, (6), 613-21.

44. Lima, T. M.; Kanunfre, C. C.; Pompeia, C.; Verlengia, R.; Curi, R., Ranking the toxicity of fatty acids on Jurkat and Raji cells by flow cytometric analysis. *Toxicol In Vitro* 2002, 16, (6), 741-7.

45. Hayakawa, T.; Kondo, T.; Shibata, T.; Kitagawa, M.; Ono, H.; Sakai, Y.; Kiriyama, S., Enzyme immunoassay for serum pancreatic lipase in the diagnosis of pancreatic diseases. *Gastroenterol Jpn* 1989, 24, (5), 556-60.

46. Denizot, Y.; Chianea, T.; Labrousse, F.; Truffinet, V.; Delage, M.; Mathonnet, M., Platelet-activating factor and human thyroid cancer. *Eur J Endocrinol* 2005, 153, (1), 31-40.

47. Dong, Q.; Patel, M.; Scott, K. F.; Graham, G. G.; Russell, P. J.; Sved, P., Oncogenic action of phospholipase A2 in prostate cancer. *Cancer Lett* 2006, 240, (1), 9-16.

48. Graff, J. R.; Konicek, B. W.; Deddens, J. A.; Chedid, M.; Hurst, B. M.; Colligan, B.; Neubauer, B. L.; Carter, H. W.; Carter, J. H., Expression of group IIa secretory phospholipase A2 increases with prostate tumor grade. *Clin Cancer Res* 2001, 7, (12), 3857-61.

49. Jiang, J.; Neubauer, B. L.; Graff, J. R.; Chedid, M.; Thomas, J. E.; Roehm, N. W.; Zhang, S.; Eckert, G. J.; Koch, M. O.; Eble, J. N.; Cheng, L., Expression of group IIA secretory phospholipase A2 is elevated in prostatic intraepithelial neoplasia and adenocarcinoma. *Am J Pathol* 2002, 160, (2), 667-71.

50. Laye, J. P.; Gill, J. H., Phospholipase A2 expression in tumours: a target for therapeutic intervention? *Drug Discov Today* 2003, 8, (15), 710-6.

51. Sved, P.; Scott, K. F.; McLeod, D.; King, N. J.; Singh, J.; Tsatralis, T.; Nikolov, B.; Boulas, J.; Nallan, L.; Gelb, M. H.; Sajinovic, M.; Graham, G. G.; Russell, P. J.; Dong, Q., Oncogenic action of secreted phospholipase A2 in prostate cancer. *Cancer Res* 2004, 64, (19), 6934-40.

52. Tribler, L.; Jensen, L. T.; Jorgensen, K.; Brunner, N.; Gelb, M. H.; Nielsen, H. J.; Jensen, S. S., Increased expression and activity of group IIA and X secretory phospholipase A2 in peritumoral versus central colon carcinoma tissue. *Anticancer Res* 2007, 27, (5A), 3179-85.

53. Yamashita, J.; Ogawa, M.; Sakai, K., Prognostic significance of three novel biologic factors in a clinical trial of adjuvant therapy for node-negative breast cancer. *Surgery* 1995, 117, (6), 601-8.

54. Yamashita, S.; Ogawa, M.; Sakamoto, K.; Abe, T.; Arakawa, H.; Yamashita, J., Elevation of serum group II phospholipase A2 levels in patients with advanced cancer. *Clin Chim Acta* 1994, 228, (2), 91-9.

55. Yamashita, S.; Yamashita, J.; Ogawa, M., Overexpression of group II phospholipase A2 in human breast cancer tissues is closely associated with their malignant potency. *Br J Cancer* 1994, 69, (6), 1166-70.

56. Yamashita, S.; Yamashita, J.; Sakamoto, K.; Inada, K.; Nakashima, Y.; Murata, K.; Saishoji, T.; Nomura, K.; Ogawa, M., Increased expression of membrane-associated phospholipase A2 shows malignant potential of human breast cancer cells. *Cancer* 1993, 71, (10), 3058-64.

57. Scheele, G. A.; Fukuoka, S.; Freedman, S. D., Role of the GP2/THP family of GPI-anchored proteins in membrane trafficking during regulated exocrine secretion. *Pancreas* 1994, 9, (2), 139-49.

58. Rampoldi, L.; Scolari, F.; Amoroso, A.; Ghiggeri, G.; Devuyst, O., The rediscovery of uromodulin (Tamm-Horsfall protein): from tubulointerstitial nephropathy to chronic kidney disease. *Kidney Int* 2011, 80, (4), 338-47.

59. Matsumoto, Y.; Imai, Y.; Sugita, Y.; Tanaka, T.; Tsujimoto, G.; Saito, H.; Oshida, T., CCDC132 is highly expressed in atopic dermatitis T cells. *Mol Med Report* 2010, 3, (1), 83-7.

60. Sekikawa, A.; Fukui, H.; Fujii, S.; Ichikawa, K.; Tomita, S.; Imura, J.; Chiba, T.; Fujimori, T., REG Ialpha protein mediates an anti-apoptotic effect of STAT3 signaling in gastric cancer cells. *Carcinogenesis* 2008, 29, (1), 76-83.

61. Usami, S.; Motoyama, S.; Koyota, S.; Wang, J.; Hayashi-Shibuya, K.; Maruyama, K.; Takahashi, N.; Saito, H.; Minamiya, Y.; Takasawa, S.; Ogawa, J.; Sugiyama, T., Regenerating gene I regulates interleukin-6 production in squamous esophageal cancer cells. *Biochem Biophys Res Commun* 2010, 392, (1), 4-8.

62. Yamagishi, H.; Fukui, H.; Sekikawa, A.; Kono, T.; Fujii, S.; Ichikawa, K.; Tomita, S.; Imura, J.; Hiraishi, H.; Chiba, T.; Fujimori, T., Expression profile of REG family proteins REG Ialpha and REG IV in advanced gastric cancer: comparison with mucin phenotype and prognostic markers. *Mod Pathol* 2009, 22, (7), 906-13.

63. Zhou, L.; Zhang, R.; Wang, L.; Shen, S.; Okamoto, H.; Sugawara, A.; Xia, L.; Wang, X.; Noguchi, N.; Yoshikawa, T.; Uruno, A.; Yao, W.; Yuan, Y., Up-regulation of REG Ialpha accelerates tumor progression in pancreatic cancer with diabetes. *Int J Cancer* 2010.

64. Adamczyk, B.; Tharmalingam, T.; Rudd, P. M., Glycans as cancer biomarkers. *Biochim Biophys Acta* 2011.

65. Cazet, A.; Julien, S.; Bobowski, M.; Burchell, J.; Delannoy, P., Tumour-associated carbohydrate antigens in breast cancer. *Breast Cancer Res* 2010, 12, (3), 204.

66. Reis, C. A.; Osorio, H.; Silva, L.; Gomes, C.; David, L., Alterations in glycosylation as biomarkers for cancer detection. *J Clin Pathol* 2010, 63, (4), 322-9.

67. Sarrats, A.; Saldova, R.; Pla, E.; Fort, E.; Harvey, D. J.; Struwe, W. B.; de Llorens, R.; Rudd, P. M.; Peracaula, R., Glycosylation of liver acute-phase proteins in pancreatic cancer and chronic pancreatitis. *Proteomics Clin Appl* 2010, 4, (4), 432-48.

CHAPTER 4

DEVELOPMENT OF WORKFLOWS AND TOOLS FOR HIGH THROUGHPUT SEMI-AUTOMATED GLYCAN MASS SPECTRAL DATA ANNOTATION

Abstract

Most membrane and secreted proteins produced by eukaryotic cells undergo post-translational modifications in the form of N-linked or O-linked glycans, giving rise to immense structural heterogeneity in mature glycopeptides. Even when cultured cells are held under tightly controlled fermentation conditions for the production of glycoprotein biologics, glycoform heterogeneity pertains and must be fully analyzed to ensure the quality and batch-to-batch consistency associated with production of biological therapeutics.

The capacity of cells to present diverse glycan structures at their surface allows them to regulate their interactions with other cells and with their environment. Human genetic disorders that affect the fidelity of glycoprotein glycan synthesis and processing result in mental retardation, skeletal and connective tissue abnormalities, anemia, multiple sclerosis, compromised immune response, muscular dystrophies, and generalized failure to thrive. Therefore, mechanisms that regulate glycan expression provide novel, broadly applicable, but largely unexplored, targets for therapeutic intervention.

The past 10-15 years have witnessed rapid growth in the interest of investigators, funding agencies, and pharmaceutical companies for characterizing glycoprotein glycan

diversity. Full and comprehensive characterization of the glycans on glycoproteins has become an essential element for drug development, quality characterization, and basic biomedical research. Manual interpretation of mass spectrometry datasets constitutes the core of most glycomics technology currently in use. However, Interpretation of up to 2000 mass spectra per biological sample consumes significant expert personnel time and reduces the number of samples that can be analyzed. This bottleneck is a major impediment blocking the expansion of glycomic analysis to a broad range of basic biomedical investigations. Progress in the field has been severely restricted by the absence of appropriate computational software tools that facilitate automated structural assignment and high throughput data analysis.

Through a collaborative partnership between PREMIER Biosoft Inc, computer scientists, experimentalists, and mass spectrometrists at the Complex Carbohydrate Research Center, tools have been developed in an effort to provide a semi-automated high throughput workflow aimed to fulfill this critical need. The development of SimGlycan® functionalities along with database development and post-processing tools described here provide innovative bioinformatic solutions.

The workflows that I have developed are conservatively expected to shift throughput from 1 sample every three months to 6 samples every week. This acceleration constitutes a paradigm shift in glycomics such that the statistical confidence afforded by increased sample number would be available to glycomics. Simply stated, our goal is to make glycomic analysis a routine, albeit technically demanding, option for a broad range of biochemical investigations.

Background And Significance

The Importance of Glycoprotein Glycosylation

The vast majority of all membrane and secreted proteins produced by eukaryotic cells are post-translationally modified by the addition of carbohydrate moieties in the form of Asn-linked or Ser/Thr-linked oligosaccharides. These N-or O-linked glycans impart immense structural heterogeneity to mature glycoproteins. As a general and well-documented rule, a single glycoprotein produced by a single cell exhibits variable usage of glycosylation sites as well as heterogeneous glycan elaboration at the glycosylation sites that are utilized. This variability in glycoprotein glycoform production has been referred to as “microheterogeneity” and is the rule for eukaryotic cells, tissues, and expression systems. Even when cultured cells are held under tightly controlled fermentation conditions for the production of glycoprotein biologics, glycoform heterogeneity pertains and must be fully analyzed to ensure the quality and batch-to-batch consistency associated with production of biological therapeutics.

Glycoform heterogeneity has been conserved and propagated across species because it provides significant advantages to cells in tissues. Cell-surface glycans mediate interactions between cells and define cellular identities within complex tissues at all stages of animal life. The capacity of cells to present diverse glycan structures at their surface allows them to regulate their interactions with other cells and with their environment. In addition, specific glycan structures frequently modulate the activities of the glycoproteins to which they are attached, adding a post-transcriptional, post-translational layer of regulation onto protein function. In some instances, glycans on specific glycoproteins serve as recognition markers that modulate cell-cell interactions among defined cell populations. In other contexts, the function of specific cell surface signaling molecules requires the elaboration of an exact glycan structure at a precise site on an appropriate glycoprotein. Human genetic disorders that affect the fidelity of glycoprotein glycan synthesis and processing result in mental retardation, skeletal and connective tissue abnormalities, anemia, multiple sclerosis, compromised immune response, muscular dystrophies, and generalized failure to thrive. Therefore, mechanisms that regulate glycan expression provide novel, broadly applicable, but largely unexplored, targets for therapeutic intervention. The full exploitation of such targets requires the development of rapid, facile, and robust tools for implementing deep and high-throughput glycan analysis.

State-Of-The-Art In Glycomic And Glycoproteomic Analysis

The past 10-15 years have witnessed rapid growth in the interest of investigators, funding agencies, and pharmaceutical companies for characterizing glycoprotein glycan diversity. With regard to basic science initiatives, the NIH (NIGMS) awarded a Glue Grant for the formation of the Consortium for Functional Glycomics (CFG) during this period, which has successfully facilitated a consolidation of efforts between Glycobiology, Glycomic, and Glycoproteomic investigators.

A major goal of the CFG has been to develop and implement optimized methods for glycan analysis in various mammalian tissues. These efforts have been extremely successful in generating large datasets of raw mass spectrometric data, but significantly less successful in developing methods for interpreting these datasets. CFG data has been primarily limited to full MS data (mostly MALDI-TOF) with occasional extension into MS/MS or MSn analysis. When such deeper analysis has been undertaken, analysis of fragmentation has been by manual interpretation. By any measure, the CFG has been successful in pushing the field forward, but this is a crucial time to consider what is needed to facilitate continued growth in glycomics and glycoproteomics beyond the era of the CFG.

Approaches for unbiased acquisition of MS and MSn data that can describe the full diversity of glycans in complex samples have been developed. These approaches, largely based on ion-trap instrumentation and automated fragmentation workflows, generate extremely large datasets. For instance, a relatively uncomplicated analysis of N-linked glycans harvested from a single animal tissue can produce over 700 separate spectra when the analysis is taken to MS2[77]. Each of these spectra currently requires manual interpretation to assess glycan presence and structural features. It is clear, however, that, in many cases, MS3 or MS4 is essential for confident structural assignment[78]. There is no tool available currently that can efficiently and practically assist analysts to sort through the mountain of data generated by standard MS instrumentation beyond MS, and especially beyond MS2.

The goals and efforts of the CFG have been primarily focused on questions related to the basic science of glycan function in complex organisms. In parallel, a broad range of drug companies, from small biotechs to large pharma, have faced the need to define the glycan diversity of their glycoprotein biologics. The FDA has, for the most part, allowed a relatively broad range of techniques to be employed for glycan characterization (HPAEC-PAD, monosaccharide composition, GC-MS, FACE, CE-LIF), as long as the methods are supported by extensive validation. The flexibility of the FDA in this regard reflects the reality that a single tool has not been aggressively developed to meet the need for comprehensive glycomic or glycoproteomic analysis. Advances in MS instrumentation, in the robustness of glycan release, and in the standardization of analytic techniques present new opportunities for proposing that MSn coupled to automated data interpretation can provide an essential and broadly applicable tool for process development, batch analysis, and quality control of drug substance in the pharmaceutical industry.

Currently Available Tools For Glycomic And Glycoproteomic Analysis

The academic community has adopted GlycoWorkbench, developed and supported by EurocarbDB as the most widely used tool for the interpretation of glycomics MS data. EurocarbDB, like the CFG in the United States, is a publicly funded endeavor to develop and implement glycomics tools and glycomics databases[65]. Also, like the CFG, EurocarbDB has ceased analytic operations. GlycoworkBench is supported in its current form through other resources, but will not be advanced or developed toward implementation of further upgrades. The GlycoWorkbench platform has served the community well, but also has some significant limitations. Among its most useful functionalities, it allows users to easily input structures of interest and generate theoretical fragmentation schema that can serve as a benchmark for interpreting real data. The graphical interfaces of GlycoworkBench were developed in consultation with practicing glycomic analysts and they have proven to be efficient and user-friendly. They provide key functionalities that are utilized throughout the workflow that we describe.

EurocarbDB had planned to make GlycoWorkbench a module that could be incorporated into workflows that would access a broader range of tools. Such workflows were envisioned to take raw MS data (and eventually MS/MS data) and provide a user-interface for assigning structures. SimGlycan® version 1.0 already provided this functionality[60a]. Thus, SimGlycan® was already at the forefront of all publicly available glycomics tools for interpreting MS data. However, SimGlycan® has significant limitations that keep it from achieving truly useful status. The purpose of our collaborative effort with Premier Biosoft was to bring SimGlycan® to the level that it can be implemented as an innovative addition to the glycomics and glycoproteomics arsenal. More recently, an alternative annotation software tool named Gelato (Glycomic Elucidation and Annotation Tool) has been developed and implemented as part of a comprehensive glycomic analysis package called the GRITS toolbox which is presented here.

Together, the gathered expertise of computer scientists, experimentalists, and mass spectrometrists provided unique opportunities to develop efficient and user-friendly computational tools that satisfy an unmet need for high-throughput and rigorous analysis of glycan modifications. The evolution of the workflow will be described in a stepwise historical fashion. The progression will follow from Version 1 to 4. Advancements, rationale, and limitations will be described as well as future directions.

Sample Preparation And Data Acquisition:

Glycans were enzymatically released from glycoproteins by PNGase F, purified over C18 columns and permethylated[77]. Permethylated glycans were dissolved in 1mM sodium hydroxide in 50% methanol, and directly infused into an LTQ-Orbitrap Discovery hybrid mass spectrometer using nano-ESI ionization. High resolution full MS spectra were captured in FT mode in the Orbitrap and fragmented by collision induced dissociation (CID) in in the ion-trap (IT mode). Manufacturer specifications reported as maximums include: Resolution 30,000 FWHM at 400 m/z, mass range of 50-2000 m/z, mass accuracy 5ppm with external calibration, dynamic range 4,000 in a single spectrum and 10,000 between spectra. RAW data was then submitted for structural assignment by SimGlycan® or GELATO and compared to manual interpretation.

Software Descriptions:

SimGlycan® by Premier Biosoft

SimGlycan® is a commercially available glycan MS/MS data analysis tool produced by PREMIER Biosoft, Palo Alto CA (www.premierbiosoft.com). SimGlycan® is a client server application that uses fragmentation fingerprinting to match experimental MS/MS data with theoretical fragments generated for structures present in a database to identify the best match. Glycan structure matches are ranked and scored by calculating how well the experimental pattern matches the theoretical pattern score. SimGlycan® calculates the glycan rank based on their proprietary matching algorithm which considers composition and branching patterns as defined by SimGlycan® as the following:

> "The composition score is a number which reflects how well the monosaccharide composition of the theoretical glycan matches that of the experimental glycan. The composition score is a function of two numbers, glycosidic percentage match and glycosidic intensity:
>
> > Glycosidic Percentage Match: The percentage of the number of theoretical glycosidic and glycosidic/glycosidic fragment peaks that have the same m/z value as that of the experimentally observed peaks compared to the total number of theoretical peaks.
> >
> > Glycosidic Intensity: The percentage of the intensity of theoretical glycosidic and glycosidic/glycosidic fragment peaks that have the same m/z value as that of the experimentally observed peaks compared to the sum of the intensities of all theoretical peaks.
>
> The branching pattern score is a number which reflects how well the topological pattern of the theoretical glycan matches that of the experimental glycan. The branching pattern score is also a function of two numbers, cross-ring percentage match and cross-ring intensity:
>
> > Cross-ring Percentage Match: Percentage of the number of theoretical cross ring and cross ring/glycosidic fragment peaks that

> have the same m/z value as the m/z values of the experimentally observed peaks compared to the total number of theoretical peaks.
> Cross-ring Intensity: Percentage of the intensity of theoretical cross ring and cross ring/glycosidic fragment peaks that have the same m/z value as the m/z values of the experimentally observed peaks compared to the sum of the intensities of all theoretical peaks.[79] "

SimGlycan® added "product diagnostic ion presence" and product ion charge state parameters that combined with existing composition and branching scores to create a newer "proximity score" in later versions.

GELATO (Glycomic Elucidation and Annotation Tool) by CCRC

GELATO is a free, semi-automated tandem MS annotation tool that was designed and implemented at the Complex Carbohydrate Research Center (CCRC) through the collaborative efforts of the author, Rene Renzinger, and Brent Weatherly.. GELATO provides a novel algorithm that rapidly matches theoretical fragmentation spectra to experimental spectra similar to SimGlycan©. The theoretical fragmentation calculations are generated from Glycoworkbench program libraries and provides a set of highly curated default glycan databases that can be used for the annotation (SweetyN and SweetyO). These databases are derived from the human curated Glycan Ontology (GlycO). In addition it is possible to use a user created database for the annotation instead. Gelato reports matches as number of peaks matched and % of the total intensity of the peaks matched in the spectrum. Gelato is capable of considering multiple charge states for each m/z reported which alleviates issues caused by incorrect charge state reporting by the instrument. Gelato is also capable of considering neutral losses including under-permethylation and loss of water. Gelato is freely available and operates within the SimianTools/GRITS toolbox. Processing time is greatly reduced compared to SimGlycan©annotation as well.

CONVERTER by CCRC

The "Converter" tool extracts key information from the annotation results of either SimGlycan® or GELATO along with information contained in the RAW data file produced by the mass spectrometer and combines this information into a sortable table that reports the following information:

- m/z, z, signal intensity, and scan number for each spectra
- The SimGlycan® rank ,% match and proximity scores and/or the GELATO match statistics peak count and % intensity coverage.
- A cartoon representation, generated by GlycoWorkbench, along with the corresponding exact mass and the associated GlycoWorkbench string.
- A set of descriptors for the assigned structure including: # Charges, # Hex, # HexNAc, # NeuAc, # NeuGc, # Fuc, # Sulfate, # Phosphate, # HexA, # Other, # Branches, # Gal-Gal disaccharide groups, # Non core Fuc, # LacNAc groups, # LacDiNAc groups, Bisecting GlcNAc present?, Core Fuc present?, Polysialic acid present?, and Glycan Type (High-Man, Bisecting, Hybrid, Complex).
- Signal intensities detected for specific diagnostic MS/MS fragment ions (e.g., m/z=660 reports a monofucosylated LacNAc terminal), providing a diagnostic for structural validation and relative comparisons between samples.
- A visual simulation of the theoretical isotopic distribution for the predicted glycan structure, which is compared to the actual isotopic distribution in order to generate a goodness-of-fit parameter. This parameter allows confidence levels to

be associated with each assignment and provides a rational basis for establishing thresholds below which MS signals are considered unreliable reporters of glycan prevalence.

MERGE by CCRC

The “Merge” tool was created so that multiple data sets that have been “converted” can be “merged” into one table. Since samples contain differing but overlapping sets of glycans, it can be difficult to easily compare glycan profiles. The Merge tool compiles the data from all samples into a single, non-redundant list which is then populated with glycan signal intensities and cartoon representations for each sample’s profile. The profiles of multiple samples are placed side-by-side for easy visual inspection and for further data processing (generation of bar graphs, hierarchical clustering, other transformations, etc.).

DatabaseBot by CCRC

DatabaseBot is a freely available stand-alone Java program developed in collaboration with Rene Ranzinger that can be used to create customized databases that improve both speed and accuracy of the annotation procedure for glycan MS/MS spectra. DatabaseBot uses the freely available database GlycomeDB as a resource for structures but also allows the user to specify a list of any structures and upload them into a new database. Using the GlycomeDB workflow, the user can set filters to only include structures of interest for a particular project. The structures can be filtered by their appearance in carbohydrate structure databases and ontologies, by the amount of fuzzy information in the structures or by predefined substructural features that should or should not be observed in candidate structures. The filters reduce the search space by eliminating redundancy and selecting only structures relevant for a particular research project, for example, glycans found only in yeast do not need to be considered in a project about human tissue. Customized databases help to improve the automated annotation of the experimental data and to minimize post-processing of SimGlycan®_or Gelato outputs. SweetyN and SweetyO are two examples of customized database created using the DatabaseBot. SweetyN consists of 895 highly curated N-linked glycan structures from the GlycO ontology and SweetyO contains # O-linked glycan structures.

SimianTools CL (command line) by CCRC

SimianTools CL is a freely available command line tool that uses the functionalities of the Converter and Merge tools combined with GlycoWorkbench to produce an extended spreadsheet format of SimGlycan® annotation results as described.

SimianTools GUI (graphical user interface) by CCRC

SimianTools GUI version 1.0 is a freely available standalone JAVA application based on the eclipse framework that was developed to replace SimianTools CL. Users can load SimGlycan® annotation results and run Converter and Merge within a user friendly visual setting that does not require knowledge of command line operation. Converted and/or Merge annotation reports are exported to excel where they can be viewed, sorted, and modified as needed by the user.

SimianTools GUI version 1.1 added the ability to manage data within the GUI, eliminating the need to have multiple applications open or to export outputs to excel. This version includes graphical user interfaces for the creation, management and display of SimGlycan®_or Gelato data annotation, both of which can be enhanced with the Converter and Merge tools. Experimental meta-data such as biological source, quantities, sample preparation, and instrument settings are collected and stored along with the primary data in a transferable file format. Version 1.1 also supports spectrum viewing and tables for converted/merged annotation results are viewable and interactive within the GUI. All primary data and possible annotations are stored but user-selected subsets can be viewed, simplifying the visual presentation and enhancing comparisons.

GRITS Toolbox GUI by CCRC

Simian Tools GUI v 1.1 was renamed the GRITS Toolbox GUI when SimianTools GUI version 1.1 framework was extended to allow the addition of 3rd party plug-ins. Other functionalities currently under development by various groups include iCRM for glycan quantification (Lance Wells), GAGID for glycosaminoglycan identification, SAGE for annotation by machine learning, and other methods for unique fragmentation pattern identification, all of which will plug into the GRITS framework.

Programming:

CCRC products including SimianTools CL, (Converter and Merge), SimianTools GUI, DatabaseBot, Gelato, and GRITS GUI were developed using java and the eclipse framework. Workflows and software tools developed over the past 3 years result from a continuum of work in progress and for purposes of discussion are divided into four versions. Each version is described followed by a discussion of advances made and remaining limitations.

Workflows:

1.0 Thermo RAW MS -> mzXML-> SimGlycan® 2.50 2009
2.0 Thermo RAW MS-> SimGlycan® version 4.02 -> SimianTools CL 2011
3.0 Thermo RAW MS-> SimGlycan® version 4.50 -> SimianTools GUI 2013
4.0 Thermo RAW MS-> Gelato and/ SimGlycan®-> GRITS Toolbox 2014

Table 4.1 Workflows

	1.0	2.0	3.0	4.0
Annotation Tool Options	SimGlycan®2.5	SimGlycan®4.02	SimGlycan®4.5 Gelato	
Throughput	low	medium	high	
#scans load	1	1500	20,000	
#scans analyzed	1	500	1000	
Database	SimGlycan®	SimGlycan®22,456 SweetieN	SimGlycan®SweetieN SweetieO	SweetieN SweetieO Others…
Output	browsing within SG framework only	Browsing SG and Csv file	SG- Browsing &Csv file Gelato-	GRITS browser & xls
Post-processing	none	SimianTools CL-command line Converter & Merge	SimianTools GUI Converter, merge, sample data storage, spectra viewer	GRITS GUI Converter, merge, sample data storage, spectra viewer

Workflow 1.0: RAW MS -> mzXML-> SimGlycan® 2.50 2009

Approximately 500-1000 RAW MS/MS spectra were collected per sample for analysis. The RAW files were converted to mzXML format by a third party converter known as SASHIMI. mzXML MS/MS spectra were both loaded into SimGlycan® 2.50 and analyzed individually. All searches were performed against the SimGlycan® database which consisted of approximately 10,000 glycans at the time. Results were ranked according to composition and branching score and viewed within the SimGlycan® application.

Workflow 2.0: Thermo RAW MS-> SimGlycan® 4.02 -> SimianTools CL 2011

Up to 1500 MS/MS spectra were loaded in native Thermo RAW format as a single batch. Searching and scoring was executed as a single job for up to 500 spectra and multiple jobs were able to be analyzed simultaneously. Searches were performed against the

SimGlycan® database containing approximately 22,500 glycans. Users could also choose to search customized databases including the highly curated SweetyN or SweetyO database created from GlycO using the DatabaseBot tool previously described. Templates of search parameters can be created and saved for future use. A proximity score is calculated based a proprietary algorithm that considers both the composition and branching score as well as the presence of important diagnostic product ions and their charge state to increase the confidence of assignments. Results can be either viewed in SimGlycan® or exported as a .csv file which reports scan number, m/z, z, rank, composition, sequence, and scores in a string. The .csv output can be converted to a sortable excel spreadsheet with added cartoon representations by the Converter and Merge Tools using SimianTools CL.

Workflow 3.0: Thermo RAW MS-> SimGlycan®4.50 -> SimianTools GUI 2013

Up to 20,000 MS/MS profiles can be loaded simultaneously to the SimGlycan®_server in native Thermo RAW format and search/scoring can be executed as a job on up to 1000 spectra at a time. The ion series appropriate to the type of fragmentation used to generate MS/MS spectra can be selected as a filter. For example, CID produces mostly glycosidic fragments and HCD/ETD produces more cross rings, therefore glycan spectra produced by CID would consider B/Y and C/Z ion fragments but not cross rings. Users can choose to use SimGlycan© database or customized databases such as SweetyN. Results can exported to Simian Tools GUI where they can be further processed by “converter” and “merge”. Users may alternatively select to annotate spectra by the GELATO algorithm within SimianTools.

Workflow 4.0 Thermo RAW MS-> GELATO and/or SimGlycan© 4.50-> GRITS Toolbox 2014

Version 4.0 has the same workflow options however, the framework has been extended to be able to accept new 3rd party plug in's as additional modules. The name was changed to GRITS toolbox.

Workflow 4.0 GELATO glycan annotation with Screenshots (Figure 4.1-4.7)

(1) A new project was created and samples were described in terms of material amount, species, adduct, sample preparation methods. Viewing preferences were set for CFG cartoon representation style. 300 RAW MS/MS spectra were loaded into GRITS for each of 5 samples. RAW data was converted to mzXML within GRITS and automatically stored in the project.

(2) Annotation parameters were selected including, derivitazation, neutral losses to consider, database selection, and allowable mass deviations with a 500 ppm mass tolerance, SweetyN database,

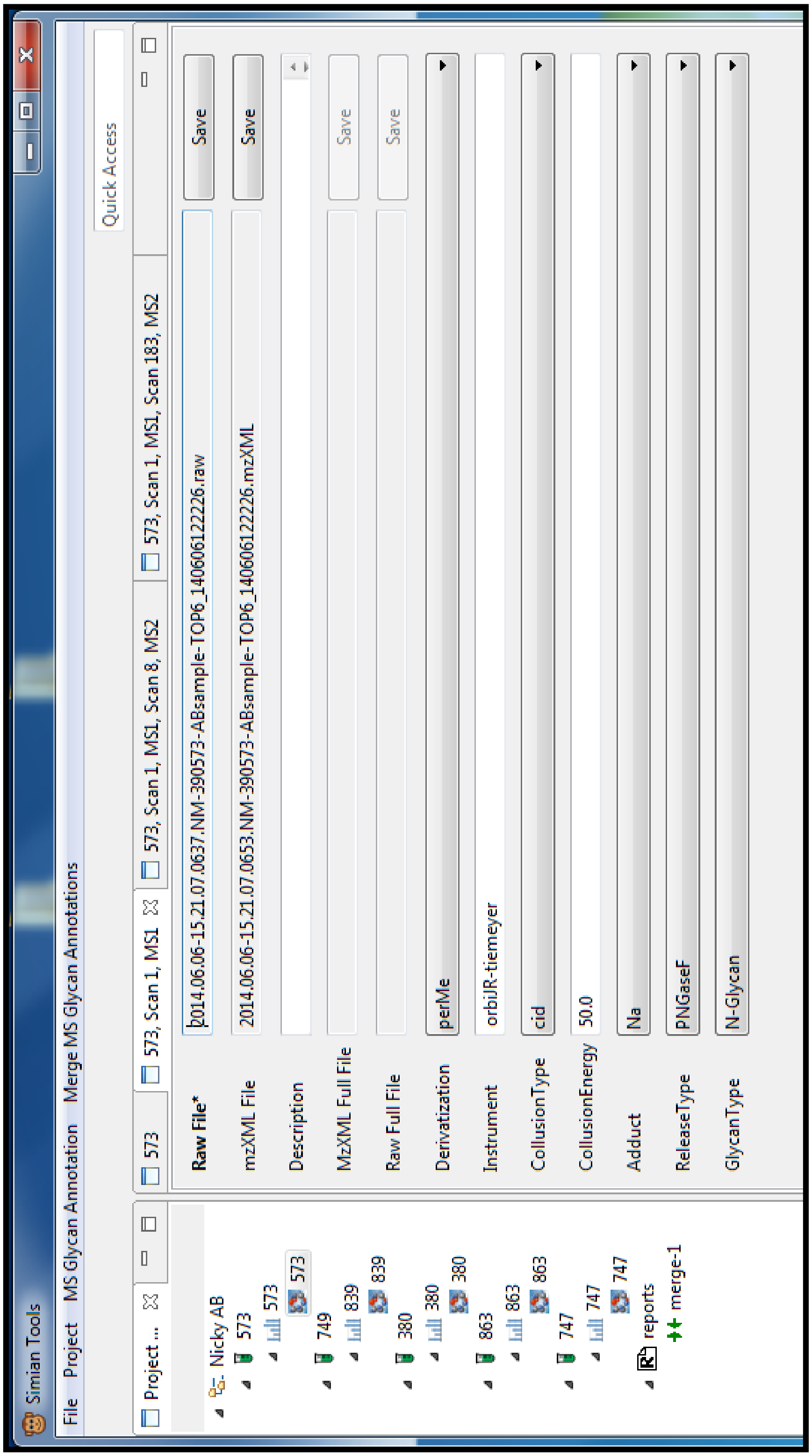
(3) Matching parameters were selected to allow up to 4 charges and fragmentation type appropriate to CID was selected. The matching was allowed to run in the background for multiple samples simultaneously and required less than five minutes to complete.

(4) Results were viewed in table format and were be compared to spectra, fragments matched, and scores.

(5) Matches were selected according to filters such as the top 3 intensity coverage or manually checked upon inspection.

(6) All five samples were merged into one table for side by side viewing.

(7) Converted and merged Results were stored and exported to excel. All original data, settings and annotations for each mass were preserved and only the view changed when annotations were manipulated. The entire workspace is portable and can be shared between researchers without the need to recreate it. The screenshots below illustrate these steps and options.



Raw File* 2014.06.06-15.21.07.0637.NM-390573-ABsample-TOP6_140606122226.raw

mzXML File 2014.06.06-15.21.07.0653.NM-390573-ABsample-TOP6_140606122226.mzXML

Description

MzXML Full File

Raw Full File

Derivatization perMe

Instrument orbiIR-tiemeyer

CollusionType cid

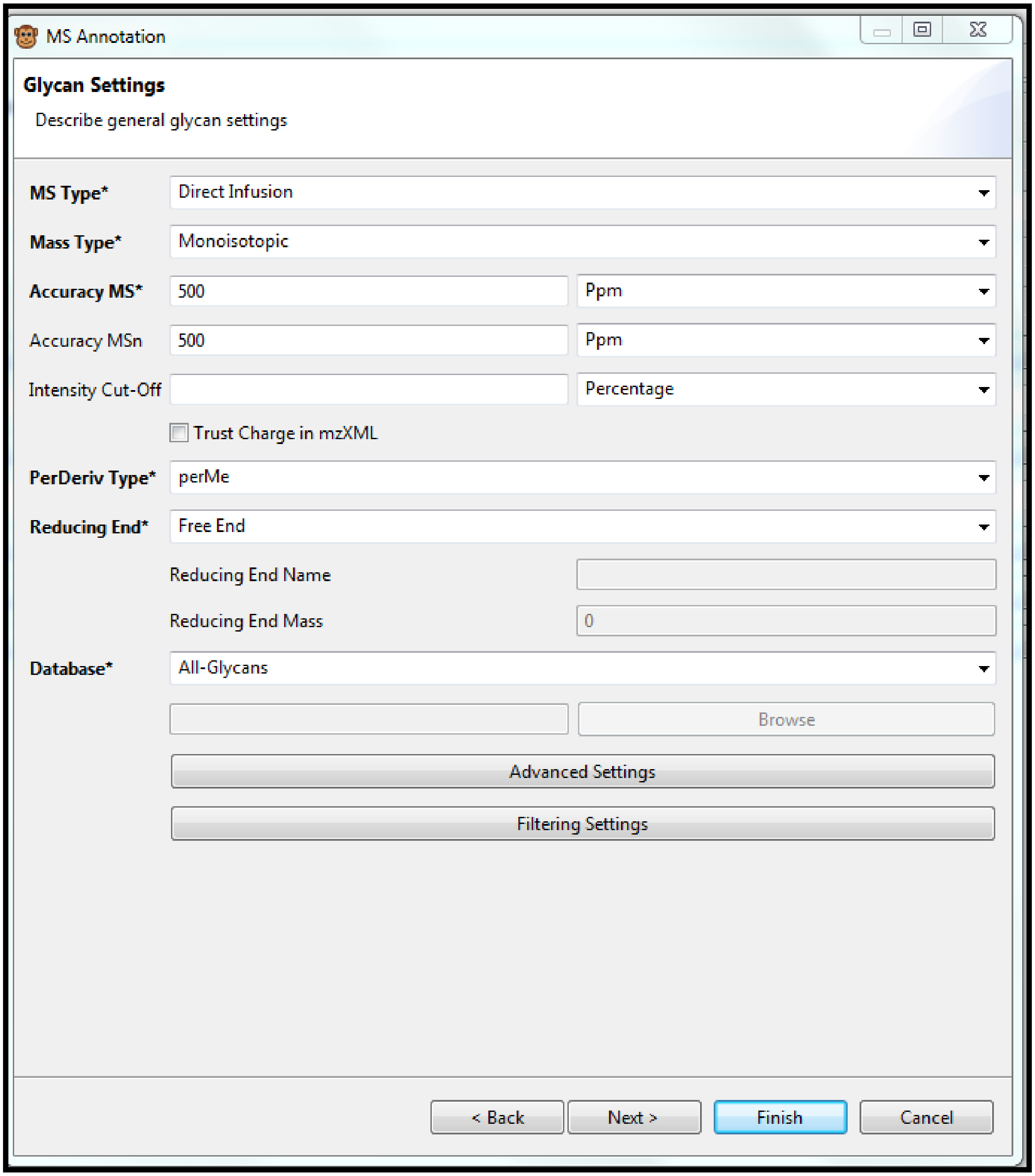
CollusionEnergy 50.0

Adduct Na

ReleaseType PNGaseF

GlycanType N-Glycan

Figure 4.1 Loading RAW data into a new project in GRITS



MS Annotation

Glycan Settings

Describe general glycan settings

MS Type* Direct Infusion

Mass Type* Monoisotopic

Accuracy MS* 500 Ppm

Accuracy MSn 500 Ppm

Intensity Cut-Off Percentage

□ Trust Charge in mzXML

PerDeriv Type* perMe

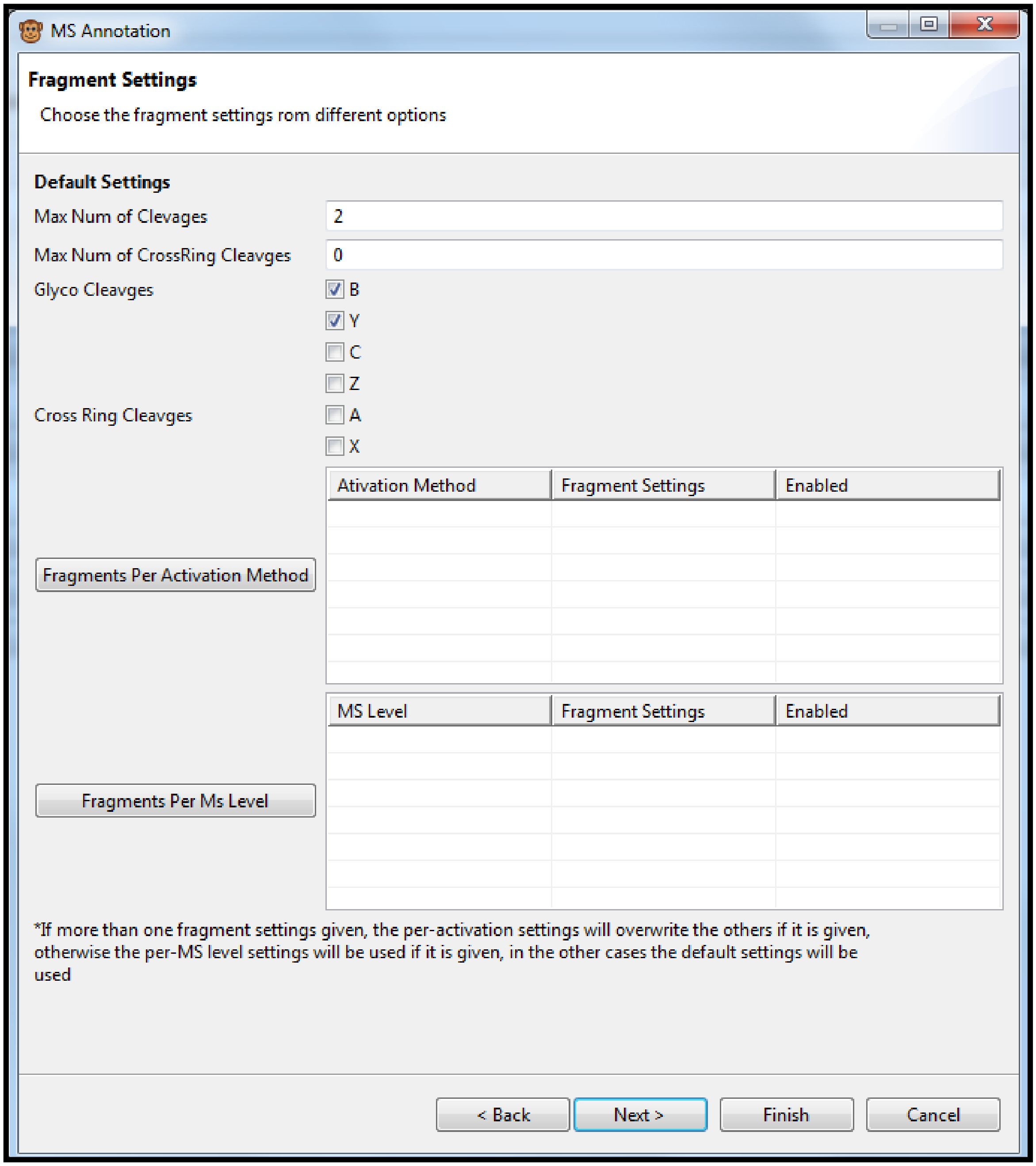
Reducing End* Free End

Reducing End Name

Reducing End Mass 0

Database* All-Glycans

Figure 4.2 Gelato annotation settings



MS Annotation

Fragment Settings

Choose the fragment settings rom different options

Default Settings

Max Num of Clevages: 2

Max Num of CrossRing Cleavges: 0

Glyco Cleavges: ☑ B ☑ Y □ C □ Z

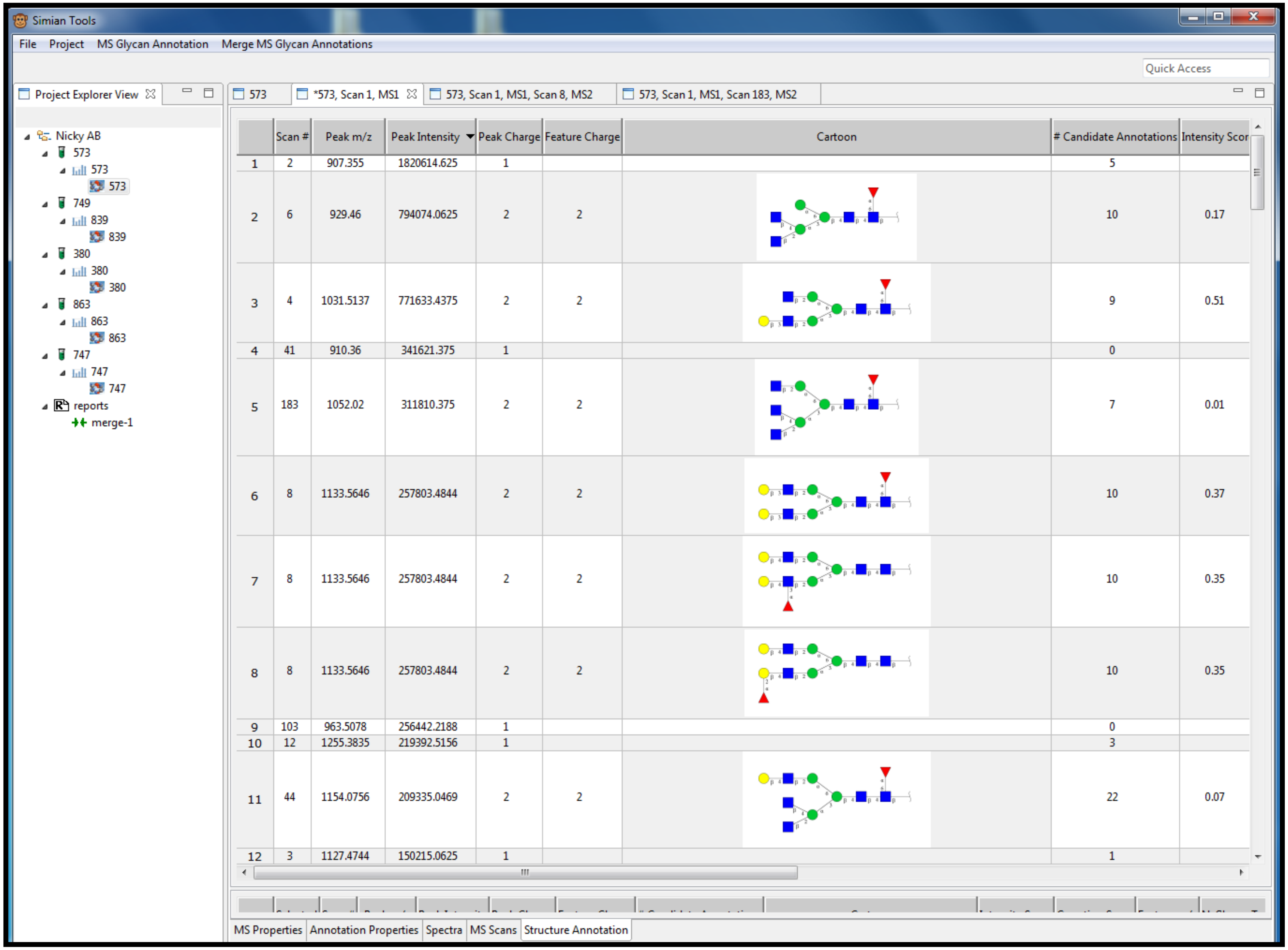
Cross Ring Cleavges: □ A □ X

Ativation Method	Fragment Settings	Enabled

MS Level	Fragment Settings	Enabled

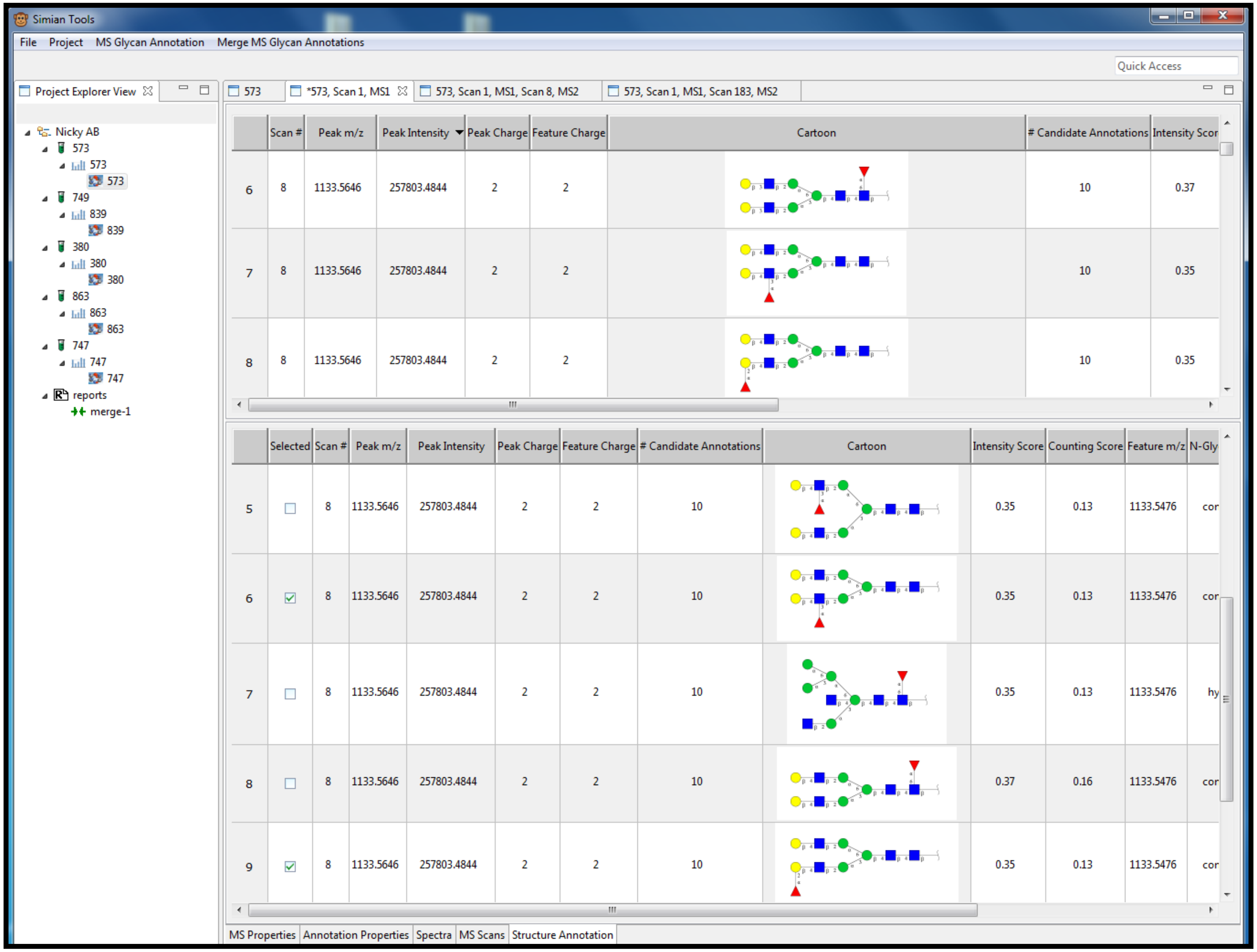
*If more than one fragment settings given, the per-activation settings will overwrite the others if it is given, otherwise the per-MS level settings will be used if it is given, in the other cases the default settings will be used

Figure 4.3 Gelato fragmentation settings



	Scan #	Peak m/z	Peak Intensity	Peak Charge	Feature Charge	Cartoon	# Candidate Annotations	Intensity Scor
1	2	907.355	1820614.625	1			5	
2	6	929.46	794074.0625	2	2		10	0.17
3	4	1031.5137	771633.4375	2	2		9	0.51
4	41	910.36	341621.375	1			0	
5	183	1052.02	311810.375	2	2		7	0.01
6	8	1133.5646	257803.4844	2	2		10	0.37
7	8	1133.5646	257803.4844	2	2		10	0.35
8	8	1133.5646	257803.4844	2	2		10	0.35
9	103	963.5078	256442.2188	1			0	
10	12	1255.3835	219392.5156	1			3	
11	44	1154.0756	209335.0469	2	2		22	0.07
12	3	1127.4744	150215.0625	1			1	

Figure 4.4 Gelato results table view

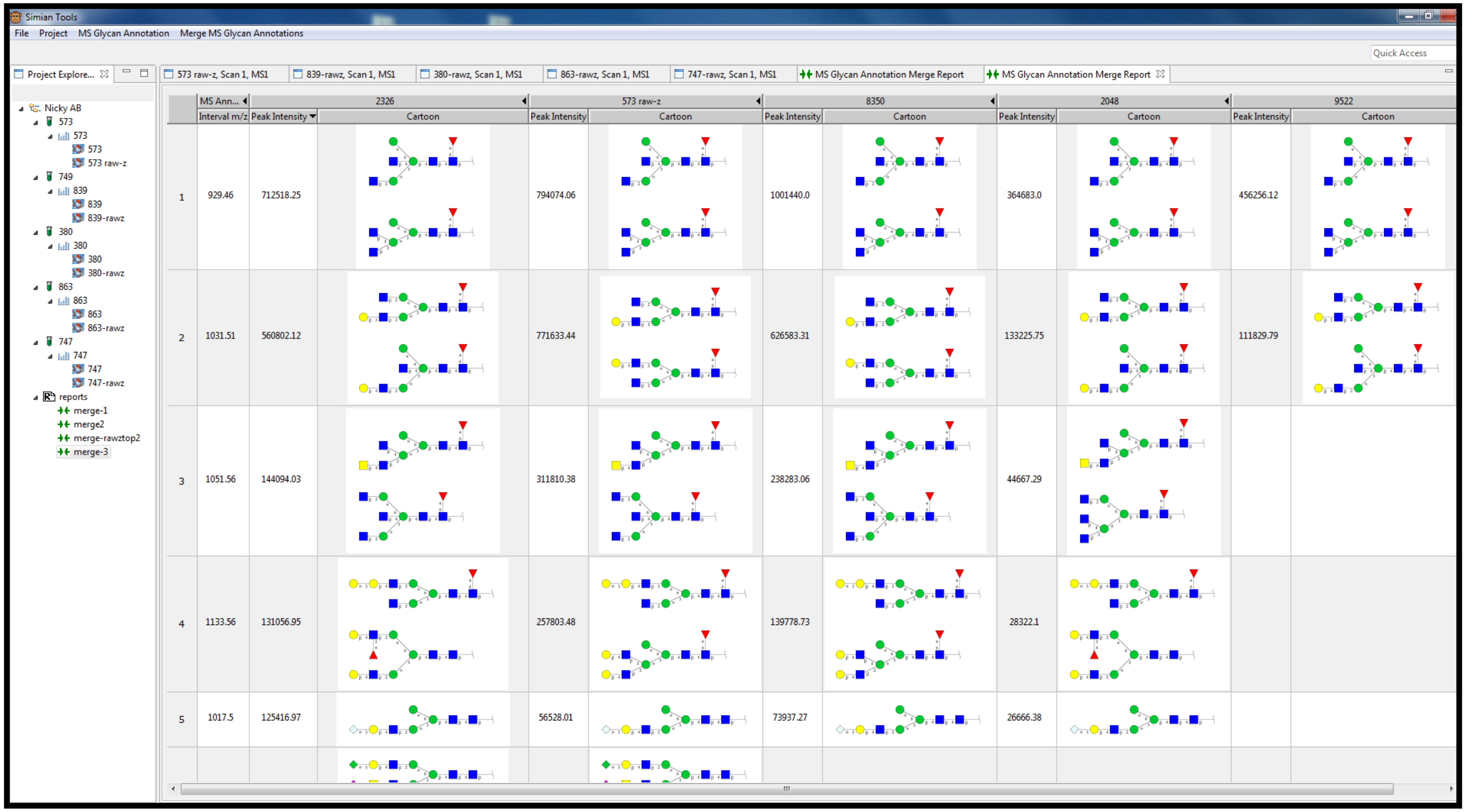


- Nicky AB
  - 573
    - 573
      - 573
  - 749
    - 839
      - 839
  - 380
    - 380
      - 380
  - 863
    - 863
      - 863
  - 747
    - 747
      - 747
  - reports
    - merge-1

	Scan #	Peak m/z	Peak Intensity	Peak Charge	Feature Charge	Cartoon	# Candidate Annotations	Intensity Scor
6	8	1133.5646	257803.4844	2	2		10	0.37
7	8	1133.5646	257803.4844	2	2		10	0.35
8	8	1133.5646	257803.4844	2	2		10	0.35

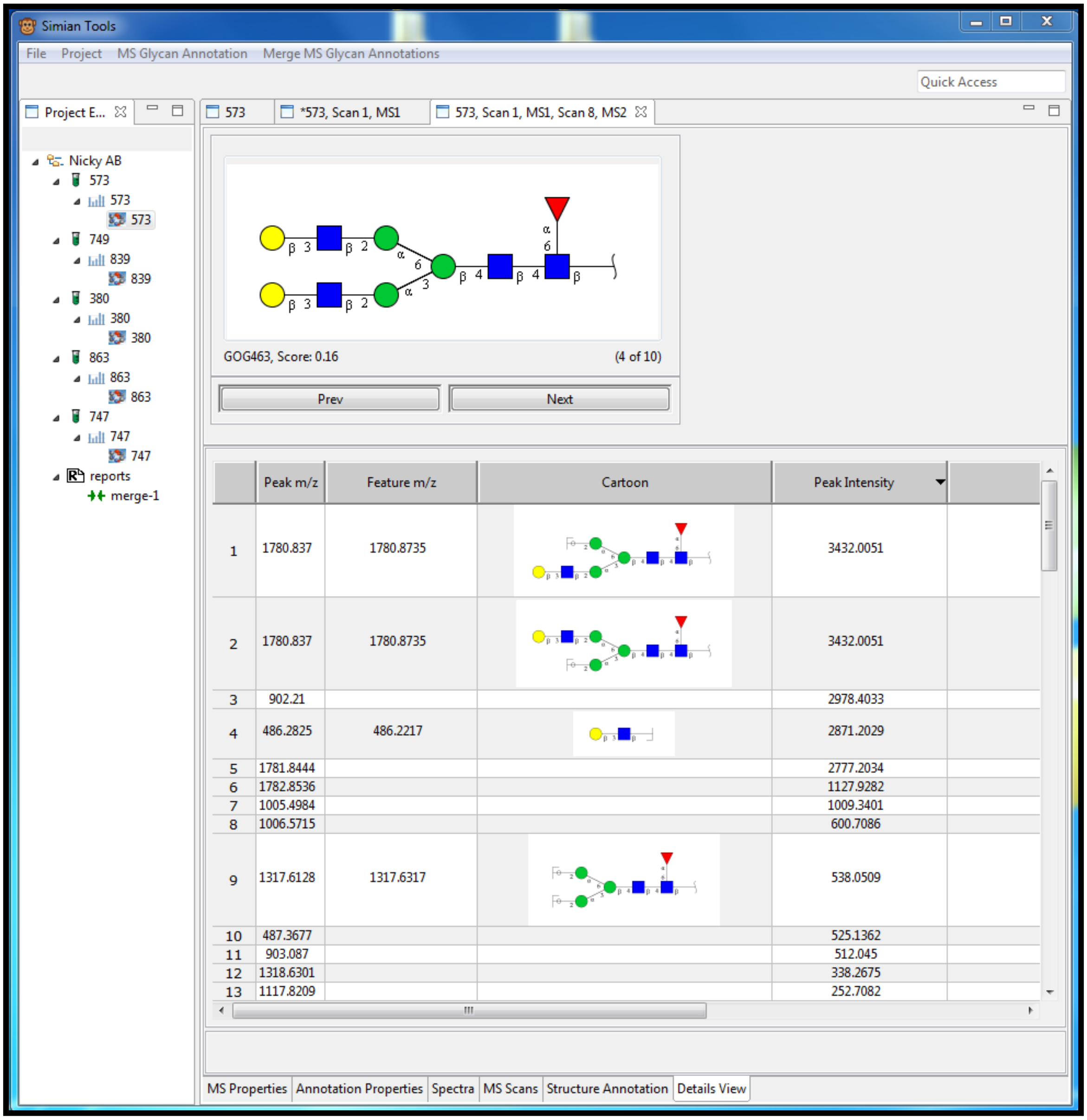
	Selected	Scan #	Peak m/z	Peak Intensity	Peak Charge	Feature Charge	# Candidate Annotations	Cartoon	Intensity Score	Counting Score	Feature m/z	N-Gly
5	☐	8	1133.5646	257803.4844	2	2	10		0.35	0.13	1133.5476	cor
6	☑	8	1133.5646	257803.4844	2	2	10		0.35	0.13	1133.5476	cor
7	☐	8	1133.5646	257803.4844	2	2	10		0.35	0.13	1133.5476	hy
8	☐	8	1133.5646	257803.4844	2	2	10		0.37	0.16	1133.5476	cor
9	☑	8	1133.5646	257803.4844	2	2	10		0.35	0.13	1133.5476	cor

Figure 4.5 Gelato results comparative view



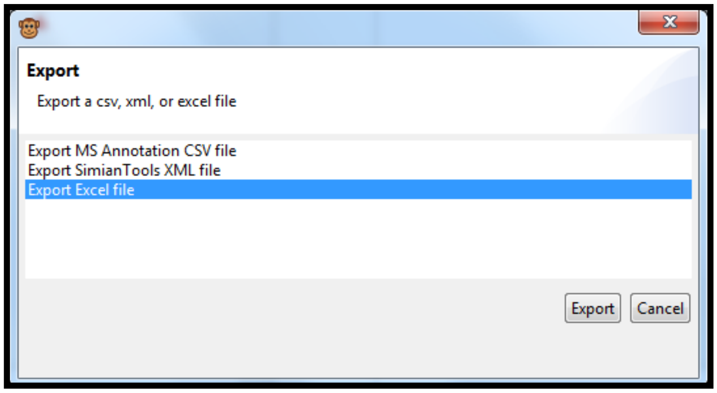
	MS Ann...	2326		573 raw-z		8350		2048		9522	
	Interval m/z	Peak Intensity	Cartoon	Peak Intensity	Cartoon	Peak Intensity	Cartoon	Peak Intensity	Cartoon	Peak Intensity	Cartoon
1	929.46	712518.25		794074.06		1001440.0		364683.0		456256.12	
2	1031.51	560802.12		771633.44		626583.31		133225.75		111829.79	
3	1051.56	144094.03		311810.38		238283.06		44667.29			
4	1133.56	131056.95		257803.48		139778.73		28322.1			
5	1017.5	125416.97		56528.01		73937.27		26666.38			

Figure 4.6 Gelato results fragment ion match view



	Peak m/z	Feature m/z	Cartoon	Peak Intensity	
1	1780.837	1780.8735		3432.0051	
2	1780.837	1780.8735		3432.0051	
3	902.21			2978.4033	
4	486.2825	486.2217		2871.2029	
5	1781.8444			2777.2034	
6	1782.8536			1127.9282	
7	1005.4984			1009.3401	
8	1006.5715			600.7086	
9	1317.6128	1317.6317		538.0509	
10	487.3677			525.1362	
11	903.087			512.045	
12	1318.6301			338.2675	
13	1117.8209			252.7082	

Figure 4.7 Gelato MERGE view



Export

Export a csv, xml, or excel file

Export MS Annotation CSV file
Export SimianTools XML file
Export Excel file

Figure 4.8 Gelato export options

Discussion:

The first requirement that was necessary to move from single analyses to high throughput analyses of thousands of spectra was the ability to handle large quantities of data in a batchable manner. This milestone was achieved rather early and without much difficulty and capacity continues to grow as computing speeds accelerate across the informatics field.

Initial fragmentation matching and ranking algorithms were simplistic and provided little in the way of objective confidence to the user. However, advances in algorithms, including the addition of the proximity score by SimGlycan®_and alternative validation by Gelato annotation, have increased confidence levels tremendously. Additional data features extracted by Converter, such as diagnostic ion intensities and isotopic distribution matching functionalities, have enhanced the user's ability to quickly validate structural assignments as well.

Searching and scoring algorithms are inherently dependent on the quality of the database which is used to search against and have posed a major road block for rapid quality annotation of glycans. However, implementation of highly curated databases, namely SweetyN, has substantially improved the quality of annotations and overall confidence in these workflows. Figure # highlights the negative consequences the presence of redundant, incomplete, and irrelevant structures have in a database search algorithm. When the same structure is entered into the database multiple times with different identifiers they are treated as unique structures and can artificially fill the top ranked matches thereby hiding true alternate structural candidates. Incorrect and incomplete structures also increase the amount of post processing time and effort required for analysis. The customized databases continue to be updated and are currently undergoing an extensive review in an effort to identify missing structures and isoform configurations. These databases will also benefit from public release, as experts will be able to create their own highly curated databases for their species or system of interest, thereby expanding workflow benefits to other research efforts.

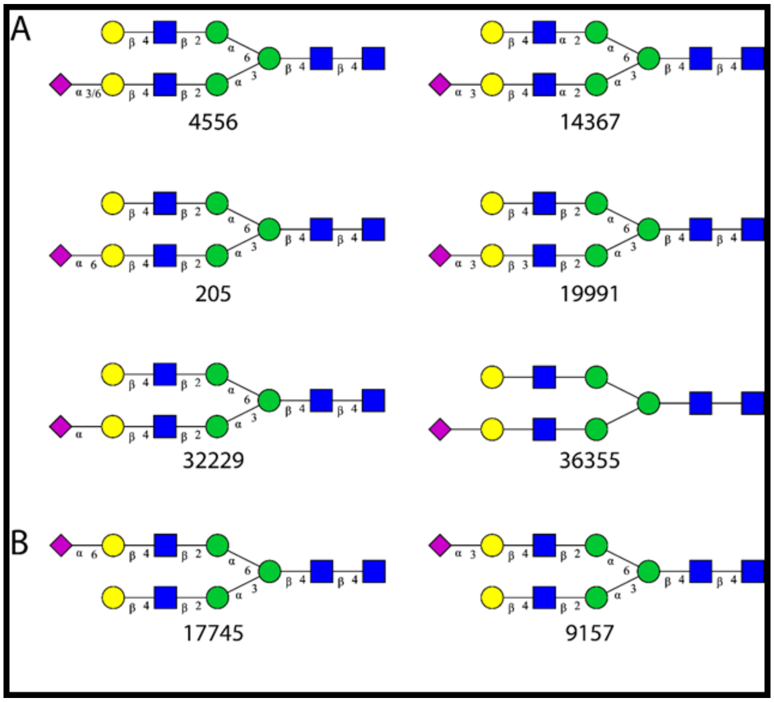


Figure 4.9 Redundant structures in public databases

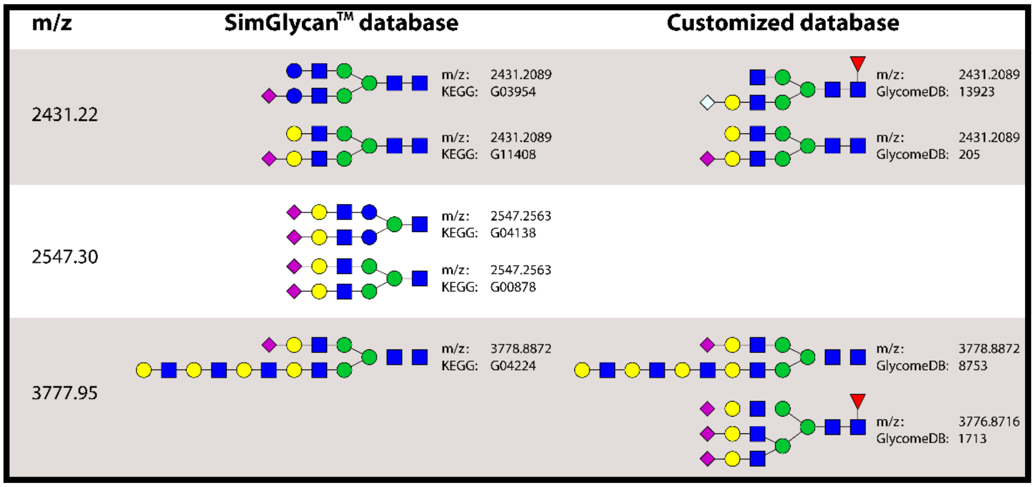


Figure 4.10 Invalid and incomplete structures hinder automated annotation of glycans

One major limitation of the type of analyses described here is the presence of mixtures of isobaric structures; ie: structures with different topologies but the same composition and therefore the same mass. This limitation does not arise from the data interpretation workflow or database quality, but from the manner in which glycans are analyzed. As described, the glycans are introduced into the mass spectrometer by direct infusion without any chromatography to separate isoforms. Therefore all structures with a given m/z are trapped together when selected for fragmentation. Two approaches are likely to resolve the issue of isobaric complexity. First, chromatographic resolution prior to MS, as is done for peptides in standard LC/MS, is possible and has been obtained for sets of non-derivatized or fluorescently tagged glycans[80]. However, separation of permethylated glycans is still a target of development. Second, targeted fragment ion methods such as the iCRM approach under development in the Wells's lab, utilize reaction monitoring approaches to distinguish fragments that differentiate isobars in an automated fashion. In theory, chromatography can be combined with intelligent reaction monitoring strategies as well.

The development of methods for chromatography of permethylated glycans will enhance the depth of discovery and the resolution of isobars. Since permethylation is necessary for maximum ionization and for in-depth structural determination, development of these approaches would provide a significant advance. LC/MS approaches will massively increase the number of spectra requiring interpretation. The workflows described here provide the foundation for this type of analyses which would have been impossible by manual interpretation methods.

The iCRM approach is a viable alternative in which product ion series unique to each isoform are identified and quantified relative to standards. This method has proven valuable for O-linked glycan analysis and should be of equal value for N-linked analysis. However appropriate fragmentation pathways for generating discriminating ions are still in development and are likely to push the limits of the duty cycle of current instrumentation. The ion series needed for identification of N-linked glycans will be more difficult to define in comparison to O-linked glycans given the large size of N-glycans (average 8-12 monosaccharides per N-glycan structure vs 3-5 per O-glycan structure) and number of structures to be defined (900 vs 300).

Great strides have been made in the journey toward high throughput glycomic analyses however much remains as well. We have built the foundation necessary to move glycomics into the fast lane. The tools we have built are works-in-progress and will continue to advance upon release to the public when additional users gain experience and suggest modifications. Additionally, the workflows provide a mechanism such that enough data can be collected to perform statistical analyses that may give rise to metrics analogous to the false discovery rate in proteomics. As results from SimGlycan®, Gelato, and manual structural assignment are collected and compared, all three techniques will benefit by identification of weaknesses and will provide opportunity for improvement as well.

References

1. Aoki-Kinoshita, K. F., An introduction to bioinformatics for glycomics research. *PLoS Comput Biol* 2008, *4* (5), e1000075.

2. Apweiler, R.; Hermjakob, H.; Sharon, N., On the Frequency of Protein Glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta* 1999, *1473* (1), 4 - 8.

3. Brooks, S., Strategies for Analysis of the Glycosylation of Proteins: Current Status and Future Perspectives. *Mol Biotechnol* 2009.

4. Varki, A.; Baum, L.; Bellis, S.; Cummings, R.; Esko, J.; Hart, G.; Linhardt, R.; Lowe, J.; McEver, R.; Srivastava, A.; Sarkar, R., Working group report: the roles of glycans in hemostasis, inflammation and vascular biology. *Glycobiology* 2008, *18* (10), 747-9.

5. Rudd, P.; Elliott, T.; Cresswell, P.; Wilson, I.; Dwek, R., Glycosylation and the immune system. *Science* 2001, *291* (5512), 2370-6.

6. Kolarich, D.; Lepenies, B.; Seeberger, P. H., Glycomics, glycoproteomics and the immune system. *Current opinion in chemical biology* 2012, *16* (1), 214-220.

7. Stevens, J.; Blixt, O.; Tumpey, T. M.; Taubenberger, J. K.; Paulson, J. C.; Wilson, I. A., Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *science* 2006, *312* (5772), 404-410.

8. (a) Roche Laboratories Inc. 2009; (b) GlaxoSmithKline. 2009.

9. Stevens, J.; Blixt, O.; Glaser, L.; Taubenberger, J. K.; Palese, P.; Paulson, J. C.; Wilson, I. A., Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *Journal of molecular biology* 2006, *355* (5), 1143-1155.

10. Kwong, P. D.; Wyatt, R.; Robinson, J.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A., Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* 1998, *393* (6686), 648-659.

11. Walker, L. M.; Phogat, S. K.; Chan-Hui, P.-Y.; Wagner, D.; Phung, P.; Goss, J. L.; Wrin, T.; Simek, M. D.; Fling, S.; Mitcham, J. L., Broad and potent neutralizing

antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 2009, *326* (5950), 285-289.

12. Karlyshev, A. V.; Ketley, J. M.; Wren, B. W., The Campylobacter jejuni glycome. *FEMS microbiology reviews* 2005, *29* (2), 377-390.

13. Varki, A.; Varki, N. M.; Borsig, L., Molecular basis of metastasis. *N Engl J Med* 2009, *360* (16), 1678-9; author reply 1679-80.

14. Marth, J.; Grewal, P., Mammalian glycosylation in immunity. *Nat Rev Immunol* 2008, *8* (11), 874-87.

15. Varki, A., *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y., 2009.

16. Consortium for Functional Glycomics. http://www.functionalglycomics.org.

17. Moremen, K. W.; Tiemeyer, M.; Nairn, A. V., Vertebrate protein glycosylation: diversity, synthesis and function. *Nature Reviews Molecular Cell Biology* 2012, *13* (7), 448-462.

18. Domon, B.; Costello, C. E., Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry. *Biochemistry* 1988, *27* (5), 1534-43.

19. Varki, N.; Varki, A., Diversity in cell surface sialic acid presentations: implications for biology and disease. *Lab Invest* 2007, *87* (9), 851-7.

20. Dwek, R., Glycobiology: Toward Understanding the Function of Sugars. *Chem Rev* 1996, *96* (2), 683-720.

21. (a) Brooks, S., Appropriate glycosylation of recombinant proteins for human use: implications of choice of expression system. *Mol Biotechnol* 2004, *28* (3), 241-55; (b) Brooks, S., Protein glycosylation in diverse cell systems: implications for modification and analysis of recombinant proteins. *Expert Rev Proteomics* 2006, *3* (3), 345-59.

22. Varki, A.; Freeze, H.; Manzi, A., Overview of glycoconjugate analysis. *Curr Protoc Protein Sci* 2001, *Chapter 12*, Unit 12.1.

23. Patel, T.; Parekh, R., Release of oligosaccharides from glycoproteins by hydrazinolysis. *Methods Enzymol* 1994, *230*, 57-66.

24. Morelle, W.; Guyétant, R.; Strecker, G., Structural analysis of oligosaccharide-alditols released by reductive beta-elimination from oviducal mucins of Rana dalmatina. *Carbohydr Res* 1998, *306* (3), 435-43.

25. Morelle, W.; Faid, V.; Chirat, F.; Michalski, J. C., Analysis of N- and O-linked glycans from glycoproteins using MALDI-TOF mass spectrometry. *Methods Mol Biol* 2009, *534*, 5-21.

26. (a) Hagglund, P.; Bunkenborg, J.; Elortza, F.; Jensen, O. N.; Roepstorff, P., A New Strategy for Identification of N-Glycosylated Proteins and Unambiguous Assignment of Their Glycosylation Sites Using HILIC Enrichment and Partial Deglycosylation. *Journal of Proteome Research* 2004, *3* (3), 556-566; (b) Morelle, W.; Michalski, J. C., Analysis of protein glycosylation by mass spectrometry. *Nat Protoc* 2007, *2* (7), 1585-602.

27. Rohrer, J., Analyzing sialic acids using high-performance anion-exchange chromatography with pulsed amperometric detection. *Anal Biochem* 2000, *283* (1), 3-9.

28. Harvey, D. J., Collision-induced fragmentation of negative ions from N-linked glycans derivatized with 2-aminobenzoic acid. *J Mass Spectrom* 2005, *40* (5), 642-53.

29. Brooks, S., Strategies for Analysis of the Glycosylation of Proteins: Current Status and Future Perspectives. *Molecular Biotechnology*.

30. Jackson, P., Polyacrylamide gel electrophoresis of reducing saccharides labeled with the fluorophore 2-aminoacridone: subpicomolar detection using an imaging system based on a cooled charge-coupled device. *Anal Biochem* 1991, *196* (2), 238-44.

31. Sharon, N.; Lis, H., History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology* 2004, *14* (11), 53R-62R.

32. Rudd, P. M.; Dwek, R. A., Rapid, sensitive sequencing of oligosaccharides from glycoproteins. *Current Opinion in Biotechnology* 1997, *8* (4), 488-497.

33. Yagi, H.; Kato, K., Multidimensional HPLC mapping method for the structural analysis of anionic N-glycans. *Trends in Glycoscience and Glycotechnology* 2009, *21* (118), 95-104.

34. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422* (6928), 198-207.

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

36. (a) Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry* 1991, *63* (24), 1193A-1203A; (b) Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* 1988, *60* (20), 2299-2301.

37. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization–principles and practice. *Mass Spectrometry Reviews* 1990, *9* (1), 37-70.

38. Cole, R. B., Electrospray ionization mass spectrometry: fundamentals, instrumentation, and applications. 1997.

39. Morelle, W.; Slomianny, M. C.; Diemer, H.; Schaeffer, C.; van Dorsselaer, A.; Michalski, J. C., Fragmentation characteristics of permethylated oligosaccharides using a matrix-assisted laser desorption/ionization two-stage time-of-flight (TOF/TOF) tandem mass spectrometer. *Rapid Commun Mass Spectrom* 2004, *18* (22), 2637-49.

40. Harvey, D. J., Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update for 2003-2004. *Mass Spectrom Rev* 2009, *28* (2), 273-361.

41. Finnigan, R. E., Quadrupole mass spectrometers. *Analytical Chemistry* 1994, *66* (19), 969A-975A.

42. Yost, R. A.; Boyd, R. K., [7] Tandem mass spectrometry: Quadrupole and hybrid instruments. *Methods in enzymology* 1990, *193*, 154-200.

43. Easterling, M. L.; Mize, T. H.; Amster, I. J., Routine part-per-million mass accuracy for high-mass ions: Space-charge effects in MALDI FT-ICR. *Analytical chemistry* 1999, *71* (3), 624-632.

44. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R., The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry* 2005, *40* (4), 430-443.
45. Scientific, T. http://www.thermo.com.

46. Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S., Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry* 2006, *78* (7), 2113-2120.

47. Harvey, D. J., Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update covering the period 1999-2000. *Mass Spectrom Rev* 2006, *25* (4), 595-662.

48. Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* 2005, *5* (4), 865-75.

49. Ceroni, A.; Dell, A.; Haslam, S. M., The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2007, *2*, 3.

50. Goldberg, D.; Bern, M.; North, S.; Haslam, S.; Dell, A., Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics* 2009, *25* (3), 365-71.
51. Ashline, D.; Singh, S.; Hanneman, A.; Reinhold, V., Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MSn. *Anal Chem* 2005, *77* (19), 6250-62.

52. Ashline, D. J.; Lapadula, A. J.; Liu, Y. H.; Lin, M.; Grace, M.; Pramanik, B.; Reinhold, V. N., Carbohydrate structural isomers analyzed by sequential mass spectrometry. *Anal Chem* 2007, *79* (10), 3830-42.

53. Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal Chem* 2005, *77* (19), 6271-9.

54. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* 2008, *4*, 222.

55. Babu, P.; North, S. J.; Jang-Lee, J.; Chalabi, S.; Mackerness, K.; Stowell, S. R.; Cummings, R. D.; Rankin, S.; Dell, A.; Haslam, S. M., Structural characterisation of neutrophil glycans by ultra sensitive mass spectrometric glycomics methodology. *Glycoconj J* 2008.

56. Lebrilla, C.; An, H., The prospects of glycan biomarkers for the diagnosis of diseases. *Mol Biosyst* 2009, *5* (1), 17-20.

57. Costello, C.; Contado-Miller, J.; Cipollo, J., A glycomics platform for the analysis of permethylated oligosaccharide alditols. *J Am Soc Mass Spectrom* 2007, *18* (10), 1799-812.

58. (a) von der Lieth, C. W.; Bohne-Lang, A.; Lohmann, K. K.; Frank, M., Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform* 2004, *5* (2), 164-78; (b) von der Lieth, C.; Lütteke, T.; Frank, M., The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim Biophys Acta* 2006, *1760* (4), 568-77.

59. Zhang, H.; Singh, S.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Anal Chem* 2005, *77* (19), 6263-70.

60. (a) Apte, A.; Meitei, N. S., Bioinformatics in glycomics: Glycan characterization with mass spectrometric data using SimGlycan™. *Functional Glycomics* 2010, 269-281; (b) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans†. *Journal of proteome research* 2008, *7* (4), 1650-1659; (c) Cooper, C.; Gasteiger, E.; Packer, N., GlycoMod--a software tool for determining glycosylation

compositions from mass spectrometric data. *Proteomics* 2001, *1* (2), 340-9; (d) Ethier, M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H., Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid communications in mass spectrometry* 2003, *17* (24), 2713-2720; (e) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A., Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res* 2007, *6* (10), 3995--4005.

61. Aoki, K.; Perlman, M.; Lim, J. M.; Cantu, R.; Wells, L.; Tiemeyer, M., Dynamic developmental elaboration of N-linked glycan complexity in the Drosophila melanogaster embryo. *Journal of Biological Chemistry* 2007, *282* (12), 9127-9142.

62. Vakhrushev, S. Y.; Dadimov, D.; Peter-KataliniÄ‡, J., Software platform for high-throughput glycomics. *Anal Chem* 2009, *81* (9), 3252--3260.

63. Goldberg, D.; Bern, M.; Li, B.; Lebrilla, C., Automatic determination of O-glycan structure from fragmentation spectra. *J Proteome Res* 2006, *5* (6), 1429-34.

64. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 2008, *7* (4), 1650-9.

65. Ceroni, A.; Dell, A.; Haslam, S., The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2007, *2*, 3.

66. Joshi, H. J.; Harrison, M. J.; Schulz, B. L.; Cooper, C. A.; Packer, N. H.; Karlsson, N. G., Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* 2004, *4* (6), 1650--1664.

67. Lohmann, K. K.; von der Lieth, C.-W., GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res* 2004, *32* (Web Server issue), W261--W266.

68. Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal Chem* 2005, *77* (19), 6271--6279.

69. Gaucher, S. P.; Morrow, J.; Leary, J. A., STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Analytical chemistry* 2000, *72* (11), 2331-2336.

70. Ethier, M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H., Application of the StrOligo algorithm for the automated

structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, *17* (24), 2713--2720.

71. Shan, B.; Ma, B.; Zhang, K.; Lajoie, G., Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *Journal of bioinformatics and computational biology* 2008, *6* (01), 77-91.

72. Tang, H.; Mechref, Y.; Novotny, M. V., Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 2005, *21 Suppl 1*, i431--i439.

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

74. Tabb, D. L.; Eng, J. K.; Yates Iii, J. R., Protein identification by SEQUEST. In *Proteome Research: Mass Spectrometry*, Springer: 2001; pp 125-142.

75. MacCoss, M. J.; Wu, C. C.; Yates, J. R., Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical chemistry* 2002, *74* (21), 5593-5599.

76. Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical chemistry* 2003, *75* (10), 2470-2477.

77. Aoki, K.; Perlman, M.; Wells, L.; Tiemeyer, M., N-linked glycan diversity in the Drosophila embryo. *Glycobiology* 2005, *15* (11), 1216-1216.

78. Reinhold, V. N.; Reinhold, B. B.; Chan, S., Carbohydrate sequence analysis by electrospray ionization-mass spectrometry. *Methods in enzymology* 1996, *271*, 377-402.

79. BioSoft, P., SimGlycan. 2009.

80. Morelle, W.; Michalski, J., Glycomics and mass spectrometry. *Curr Pharm Des* 2005, *11* (20), 2615-45.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

A dense and complex layer of glycans coat every living cell on Earth and therefore any interaction with the cell must involve glycans.   Glycans are involved in a wide range of cellular processes as part of normal physiology, development and cell signaling and are involved in every disease known to man. Glycans are the most structurally diverse and ubiquitous protein modification known and the information content encoded by them is incomprehensible. Structural diversity arising from the non-linear, non-template driven nature in which monosaccharide building blocks get linked together as glycans affords endless possibilities to affect the structure and function of carrier proteins. Glycans biosynthesis relies on environmental factors such as pH and nutrient availability and ultimately provides information to complete the link between our genome and our expressed traits, or phenotype. Deciphering the glycome will not only expand our fundamental understanding of human health and disease but the field of biology as a whole and will contribute to the development of new therapeutics as well.

Over a half a million deaths from cancer occur each year in the U.S. with costs estimated in the hundred billions, patients and society suffer as a whole. Early treatment saves both lives and money. Given that abnormal glycosylation is a universal feature of all cancer cells, and many existing clinical tests rely on glycoprotein detection, studies to further identify glycan based biomarkers as presented in Chapter 3 are well justified. One of the most important finding that came out of the glycomic side of this work was the unexpected clustering of patients into discrete subgroups that were enriched with either sialylation or fucosylation or a mix of the two. This finding was particularly insightful when considering human data sets since discrimination between cancer and non-cancer was only possible within each subset but not within the set as a whole. While it is clear that glycans are promising biomarkers and may provide targets for therapy much is still to be learned. And without a better understanding of the glycome, progress toward detecting and battling cancer along with all diseases will not be possible.

Glycomic analyses are challenging both experimentally and intellectually and currently practiced by a limited number of highly specialized laboratories. Existing tools are limited and inadequate to expand the field to non-specialists. The tools and workflows developed and presented in Chapter 4 provide a paradigm shift for glycomic analyses. The acceleration and ease of data interpretation afforded by SimianTools/GRITS will propel glycomic analysis by specialist and non-specialist investigators to the next level. Ultimately, this work will be a real option for a broad range of biomedical investigators.

While SimianTools/GRITS is an efficient and comprehensive tool for glycomic analyses, it is just a starting point. Much more is needed for the field and many aspects are undergoing rapid development now. The databases created by GlycO are the foundation of these tools are being expanded daily and will undergo a massive interrogation and expansion in 2014. Modules for improved quantification are needed and currently under development. Advanced scoring mechanisms and Isobaric separation of glycan structures are especially needed. Accomplishing these goals will provide a next big step in the progression toward truly high throughput automated glycomics.