Decoding Structure-function Relationships of Glycans

By
Nathan Wilson Stebbins
B.S., Biochemistry, Cellular and Molecular Biology
University of Tennessee, 2011

Submitted to the Department of Biological Engineering
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biological Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January 2017 [February 2017]

Signature of Author Signature redacted
Nathan Stebbins
Department of Biological Engineering
January 2017

Certified by Signature redacted
Ram Sasisekharan, Ph.D.
Alfred H. Caspary Professor of Biological Engineering
Thesis Supervisor

Accepted by Signature redacted
Mark Bathe, Ph.D.
Professor, Department of Biological Engineering
Chair, Graduate Program

This doctoral thesis has been examined by a committee of the Department of Biological Engineering

Katharina Ribbeck, Ph.D.
Eugene Bell Career Development Professor of Tissue Engineering
Department of Biological Engineering
Thesis Committee Chair

Signature redacted

Ram Sasisekharan, Ph.D.
Alfred H. Caspary Professor of Biological Engineering
Thesis Supervisor

David Berry, M.D. Ph.D
General Partner
Flagship Ventures

Decoding structure-function relationships of glycans

By

Nathan Stebbins

Submitted to the Department of Biological Engineering
on January, 2017 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Glycans are an important class of biological molecules that regulate a variety of physiological processes such as signal transduction, tissue development and microbial pathogenesis. However, due to the structural complexity of glycans and the unique intricacies of glycan-protein interactions, elucidating glycan structure–function relationships is challenging. Thus, uncovering the biological function of glycans requires an integrated approach, incorporating structural analysis of glycans, and glycan-proteins interactions with functional analysis. In this thesis, I develop new tools and implement integrated approaches to study glycans and glycan-binding proteins (GBPs). I apply these approaches to study glycans and GBPs in two areas: i) the role of hemagglutinin-glycan receptor specificity in human adaptation and pathogenesis of influenza and ii) the function of glycan regulation of cell-microenvironment interaction in cancer progression.

Section 1: Influenza poses a significant public health threat and there is a constant looming threat of a pandemic. Pandemic viruses emerge when avian viruses acquire mutations that enable human adaptation, leading to infection of an antigenically naïve host. Influenza Hemagglutinin (HA), and HA-glycan receptor interactions, play a central role in host tropism, transmissibility, and immune recognition. In section one, I develop and apply an integrated approach comprised of structural modeling, inter-amino acid network analysis, biochemical assays, and bioinformatics tools to study the hemagglutinin-glycan interaction and, in some cases, HA's antigenic properties. Using this approach, we i) identify the structural determinants required, and potential mutational paths, for H5N1 to quantitatively switch it's binding specificity to human glycans receptors, ii) identify the mutations that enable the 2013 outbreak H7N9 HA to improve binding to human glycan receptors in the upper respiratory tract, iii) uncover H3N2 strains that are currently circulating in birds and swine that possess features of a virus that could potentially re-emerge and cause a pandemic, and iv) characterize the glycan binding specificity of a novel 2011 Seal H3N8 HA. The approaches implemented here and the findings of these studies provide a framework for improved surveillance of influenza viruses circulating in non-human hosts that pose a pandemic threat.

Section 2: Glycans are abundant on the cell surface, and at the cell-ECM interface where they mediate interactions between cells and their microenvironment. Despite this, the function of glycans in cancer progression remains largely understudied. Here, I develop an integrated approach to characterize the cell surface glycome, including N-linked, O-linked glycans, and HSGAGs. This approach integrates glycogene expression data, analytical tools, and glycan binding protein reagents. I demonstrate that this platform enables rapid and efficient characterization of the N- and O-linked glycome in a model cell system, representing metastatic versus non-metastatic cancer cells. Next, I apply this integrated approach to uncover new roles of glycans. I study the role that HSGAGs play in regulating cancer stem cell (CSC) activity in breast cancer. Here, we report that SULF1, an HSGAG modifying enzyme, is required for efficient tumor initiation, growth and metastasis of CSCs. Furthermore, we identify a putative mechanism by which SULF1 regulates interactions between CSCs and their microenvironment. The approaches implemented here and the finding of these studies have important implications for the development of cancer therapeutics.

Overall, this thesis provides important tools, approaches and insights to enable and improve the study of glycans and glycan binding proteins. Together the work here provides a framework for decoding structure-function relationship of glycans.

Thesis Supervisor: Ram Sasisekharan, Ph.D.
Title: Professor of Alfred H. Caspary Professor of Biological Engineering

Acknowledgements

First and foremost, I would like to thank my thesis advisor Ram Sasisekharan. The completion of this thesis work would not have been possible without his constant support, positivity and guidance. Ram, from our very first meeting MIT, you instilled such great confidence in me. You gave me a place to learn and the freedom to grow. Your outstanding mentorship and steadfast focus on impact shaped how I think about problems and create solutions. From studying bird flu's path to pandemicness, to investigating sugars on cancer cells, to making drugs from dust & bugs, to creating MITx courses and so much more, you provided so many unique and meaningful opportunities that have helped me grow into the scientist I am today. Furthermore, your teaching out of the lab is something I greatly appreciate. Our countless candid conversations and your 'think-big' approach on things like science, industry, education, and public health truly changed the way I see the world and gave me the confidence to shape it into a better place. Again, I thank you for the genuine bond we've fostered over the years; you have had an immeasurable impact on my life and future endeavors, and I look forward to the day that we get to work together again.

I would also like to thank the other members of my thesis committee, Katharina Ribbeck and David Berry. Your feedback and guidance was essential to the completion of this thesis work. Your diverse perspectives and expertise were invaluable to this process and I am grateful to have been able to share my work with you over the years. Additionally, I would like to acknowledge my funding sources, The NIEHS Toxicology Training Grant and The Lemelson Presidential Fellowship, for their generous financial support during my PhD.

I would like to thank all of the members of the Sasisekharan lab that I encountered during my tenure. In particular, I would like to thank several people that were key to the completion of this work. To Rahul Raman, Karthik Viswanathan, Akila Jayaraman and Kannan Tharakaraman, thank you all for your mentorship, training, and guidance. You all taught me everything I know about glycobiology and influenza as well as the ins and outs of working in the Sasisekharan lab. To Andrew Hatas, your scientific and technical insight was instrumental in the completion of my work and we are

fortunate to have such a stellar lab manager. To Devin Quinlan, you have been a fantastic colleague and friend. Your energy, humor and support have made this journey unforgettable. To Ada Ziolkowski, you are a constant source of positivity and productivity for the lab, but also for me personally. Thank you for all your help and support during my time in lab. To the rest of the Sasisekharan lab members, Vidya Subramanian, Troy Rurak and all former lab members and visiting scientists, thank you for making the early mornings and long nights in lab worthwhile. From our day-to-day discovery, to our mischievous lab antics, you all made our space more of a home than a lab for the last 5 years, so thank you for that.

To make an impact in science, collaboration is key and I've had the pleasure of working with a wide collection of outstanding individuals from here in Cambridge, to Thailand and Germany. Thank you to all my collaborators: Bob Weinberg, Christine Chaffer, Erika von Mutius, Eric Ma, Islam Hussein, and John Rundstadler. It was a pleasure to work with such enthusiastic and intelligent individuals. You all taught me a great deal and shared your expertise to give me a broader understanding of the scientific world.

In addition to providing students with world-class academic training, MIT is deeply invested in the personal and professional development of its community. I am thankful to MIT for the support and freedom to create new opportunities for professional growth of its students (vis-à-vis MIT Biotech Group & 20.930J). To Dean Ian Waitz, Donna Savicki, Suzanne Glassburn: Thank you for all your support, resources, and your shared commitment to our vision. To Doug Lauffenburger, thank you for welcoming me to the BE community at MIT and providing me with a platform to give back to the department that gave so much to me. To Raven Reddy: I could not have accomplished any of this without you. You are an inspiring friend and I hope we find a reason to work together again. Lastly, thank you to the MIT Biotech Group leadership (current and former) and member base. This group is a huge source of energy and creativity at the MIT and I'm excited to see its future growth.

To everyone from the BE department and the greater MIT community that I've come into contact with over the years, thank you for being part of my experience. I am privileged to be a member of such a devoted and curious community that is dedicated to solving big problems and making a big impact. It was an honor to be a part of a community whose affiliation is not defined by culture, geography, or background, but by a singular aim to make the world a better place. You are creative and bold, and you've taught me to be daring in my approach to change this world.

Another huge thank you goes out to my amazing friends. To the boys: James Weiss, Brian Bonk, Raja Srinivas, Vyas Ramanan, Andrew Warren, Rob and Erika Kimmerling, Alec Nielson and Baris Sevinc. I am so lucky to have such an inspiring groups of friends. Thank you for being there through it all and managing to make enough memories with me to last a lifetime. To my BE classmates: Thomas Segall-Shapiro, Daniel Rothenberg, Erica Ma, Allison Claas, Gabi Pregernig, Tony Kulesa, Jacob Barrajo, Fei Chen and many more not mentioned here, thank you for your friendship over the years. I look forward to see where your future takes you. To the rest of my MIT crew: Andreas Miller, Dan Congreve, Markus Eizinger, Oliver Dodd, Christa Milley and to my UTK friends that have been there from the start, Brandon Birckhead, Ryan, Amy and Zelda Rickels, Mike Jungwirth, Marina Mikhailovna, and Tucker Netherton, your friendship means the world to me. Thank you for everything; I couldn't imagine a better group of friends to call my own.

It would not have been possible for me to be where I am without the support and encouragement of my family, near and far. I would like to thank my Mom and Dad for their unconditional love and support. You were always there to help or lend an ear when I needed it. Thank you to my sister Emily and all my Grandparents for their love and encouragement.

Finally, to Sylvia, thank you so much for being there through this process. I could not have done this without your constant love and support. Words truly cannot express how lucky I feel to have you in my life and I am so excited to see where the future takes us. I love you.

Table of Contents

Chapter 1 : Introduction

Motivation and Overview

Fundamentals of glycobiology

After sequencing the human genome, the most puzzling finding was the small number of protein encoding genes identified (~30,000). Given the phenotypic complexity of *Homo sapiens* compared to *D. melanogaster*, *C. elegans*, or *S. cereivisae*, initial estimates of the human genome size were 2-5 times greater than what was ultimately reported by Venter et al. [1]. This finding led authors to consider the importance of non-genetically encoded mechanisms in generating the functional diversity of proteins, namely, post-translational modifications (PTMs) [1]. The modified central dogma now includes PTMs as key mediators of cellular phenotype and, in this post-genomics era, PTMs have become a central focus of study. Glycosylation, defined as the covalent attachment of a carbohydrate to a glycoconjugate carrier, is the most abundant and diverse PTM observed in nature [2]. The attachment of a glycan moiety encodes a vast array of information associated with cellular function. In mammals, glycans are found attached to cell surface proteins and lipids, although they may also exist as free reducing sugars (e.g. human milk glycans), unconjugated in the extracellular matrix (ECM) (e.g. Hyaluronic acid), or attached to intracellular proteins (e.g. O-GlcNAcylation) [2]. Structurally, glycans are either branched or linear, and further classified by the nature of linkage to the glycoconjugate. Branched glycans can be Asn-linked (N-linked) or Ser/Thr-linked (O-linked) to a glycoprotein, or attached to a glycolipid [2,3]. In contrast, linear glycans, that are predominantly glycosaminoglycan (GAGs), are Ser/Thr-linked to a proteoglycan core [2,4,5] (Figure 1.1).

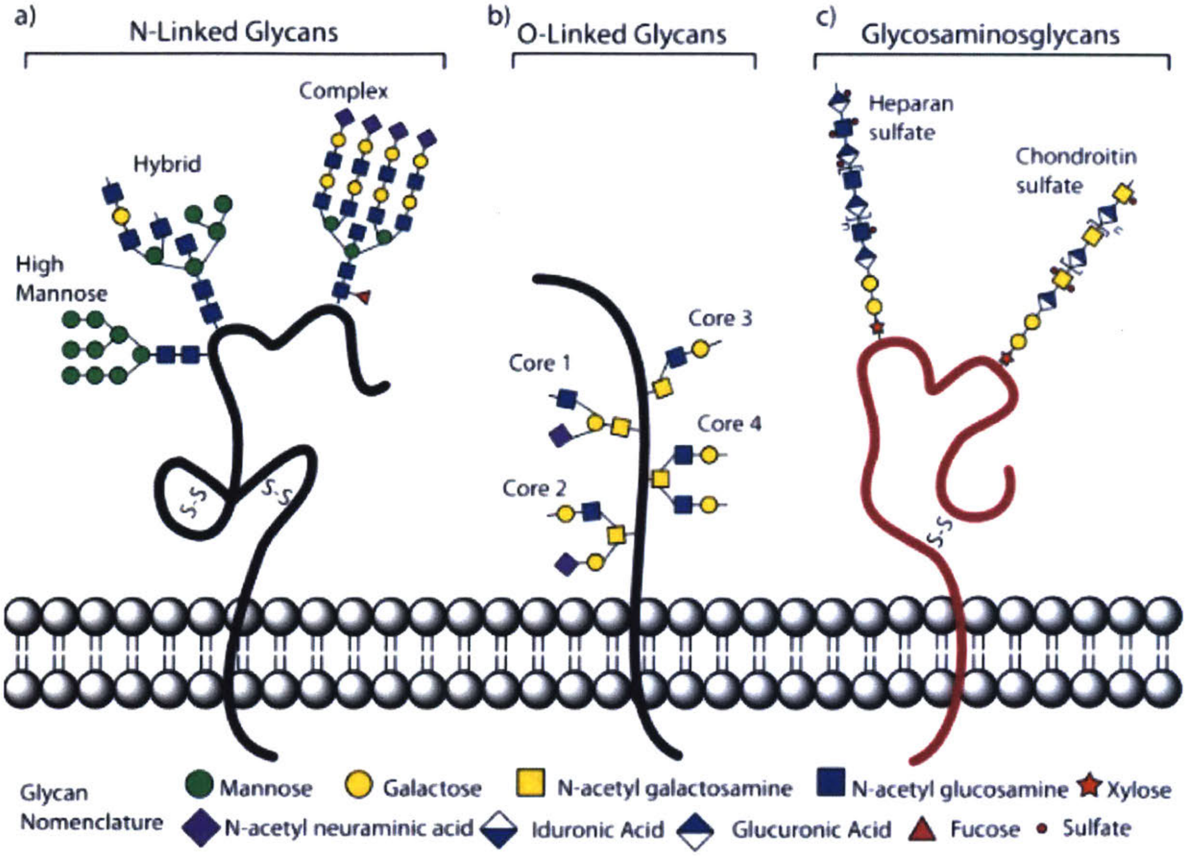


Figure 1.1 **Structural diversity of glycans. Glycans are classified based on topology and linkage the glycoconjugate.** They can be branched (N-linked and O-linked glycans) or linear (Glycosaminoglycans). Glycans structures are depicted using CFG cartoon notation. (a) N-linked glycans are characterized by the carbohydrate attachment to a protein via the asparagine of the polypeptide sequon, N-X-S/T, where X is any amino acid other than proline. All N-glycans share a common chitobiose core ($Man_3GlcNAc_2$) and are categorized into three subtypes based on the nature of the extension of the core: high-mannose, complex, and hybrid. (b) Mucin-type O-linked glycans are characterized by the covalent linkage of GalNAc to a serine or threonine, and are classified into 8 different core structures (core 1-4 depicted). Unlike N-linked glycans, O-glycans don't have a single peptide consensus sequence and several GalNAc polypeptide transferase (pp-GalNAc-Ts) enzymes utilize unique motifs for O-GalNAcylation. (c) Glycosaminoglycan's (GAGs) are linear polysaccharides composed of repeating hexosamine-uronic acid disaccharide unit and are O-linked to a proteoglycan core. Several classes of GAGs exist and are classified based on their disaccharide composition; heparin/heparan sulfate (IdoA/GlcA-GlcN), and chondroitin sulfate (GlcA-GalNAc) are depicted here. Adapted from [76]

Glycans interact with many components of the extracellular milieu including ECM proteins, growth factors, receptors, enzymes, other cells, and pathogens [2,4]. Through these interactions glycans regulate diverse processes including cell growth [6,7], development [8,9], morphogenesis [10], cell adhesion [11,12], cell-cell interaction

[13], immunity [2,14], and host-pathogen interactions [15,16]. The glycome is highly dynamic and is remodeled biosynthetically or post synthetically during various biological processes or changes in cellular state [7,17]. As such, changes in composition or abundance of glycans are observed during development, tumorigenesis and metastatic progression [18–21], auto-inflammatory disease [22], and in the immune evasion of pathogens [23].

In addition to the regulation of cell physiology in both normal and disease states, glycans exert their function at a molecular level. Glycosylation influences the structural properties of proteins, regulating features such as protein folding, solubility, stability, immunogenicity, and clearance [24–26]. These features are of interest to pharmaceutical industries and have been studied extensively in the context of developing effective biologic therapeutics. For example, monoclonal antibodies contain N-glycosylation sites in the Fc region. Interestingly, glycosylation in the Fc domain of IgG_1 is required for antibody dependent cell-mediated cytotoxicity (ADCC) or complement dependent cytotoxicity (CDC) [26,27]. Furthermore, alterations in the structure or composition of Fc glycans can enhance or diminish ADCC/CDC (e.g. hypersialylation or increased core fucosylation results in decreased ADCC/CDC [25]).

Challenges with decoding structure-function relationships of complex glycans

Glycosylation plays an integral role in a diverse array of biological and disease processes, yet decoding structure-function relationships of complex carbohydrates has proved difficult (Table 1.1). Key aspects of glycan chemical properties and fundamental modes of action suggest a need for a unique experimental toolkit and a systems level approach for determining biological function. First, glycans possess several unique structural elements (i.e. branching, anomericity, varied monosaccharide linkages) that give rise to diverse chemical structures and high degree of isomerism [28–30]. These features pose significant barriers pertaining to the development of analytical techniques capable of elucidating fine chemical structure of glycans. Next, glycan biosynthesis is a non-template driven, non-proofreading process, and involves the concerted activity of several hundred enzymes, including glycosyltransferases, glycosidases, and enzymes involved in sugar-nucleotide biosynthesis and transport [2,4,30].

Their biosynthetic complexity complicates the use of classical molecular biology or functional genomic strategies for interrogating glycan structure-function relationships. Furthermore, the lack of a template or proof reading mechanism leads to the expression of a heterogeneous, often related, set of glycan structures, and results in microheterogeneity and variation in site occupancy at the level of the glycoconjugate. Considering these attributes, cell surface glycans are presented as a heterogeneous ensemble of complex structures that modulate cellular function. The functional information contained in this glycan ensemble is interpreted through interactions with glycan binding proteins (GBPs). While protein-protein interactions are high affinity (10^{-9} - 10^{-12} M) 'digital' modulators of function, glycan-protein interactions are generally low affinity (10^{-3} - 10^{-6} M), and affect function in an 'analog' fashion [30]. Proteins that interact with the glycan ensemble may engage multiple discrete glycan structural epitopes. High affinity and specificity, which lead to signaling, are achieved through multivalent interactions (avidity) [31]. As such, glycan sequence alone is not the sole determinant of glycan-protein interactions. Features like glycan density or localization play important roles in determining activity. Thus, understanding the biochemical basis for glycan-protein interaction is challenging and conventional biochemical methods cannot accurately capture such interactions. Therefore, in order to elucidate the structure-function relationships of complex carbohydrates, we require a unique approach which leverages the integration of a structural analysis of glycans, and glycan-protein interactions with functional analysis across genetic, cellular, and organismal levels (Figure 1.2).

Key challenge	Features	Impact on study of glycans
Glycan biosynthesis	Non-template-driven process, unlike DNA/RNA and protein	Replication- or translation-like 'rules' cannot be easily applied; no direct methods to amplify glycans, unlike DNA (PCR) and protein (recombinant expression)
	Limited availability of glycans from natural sources (e.g., cells, tissues)	Without amplification tools, analytical and functional methods often require high sensitivity
	Tissue-, developmental-, and metabolic-dependent expression of glycan biosynthetic machinery	Glycan structure is sensitive to cellular conditions, tissue type, and developmental stages
	Lack of proofreading in glycan biosynthetic process	Increases structural diversity of glycans to be analyzed
Glycan structural complexity and heterogeneity	Presence of isomers and different anomeric configurations	Properties generally not present in DNA and proteins; challenges structural characterization by single method
	Microheterogeneity – a range of glycan structures (length, composition, branching) found at any given glycosylation site on a glycoprotein	Highly similar physicochemical properties of glycan microheterogeneities challenges their characterization
	Branching	Unambiguous designation of branches and their locations challenged by analytical approaches
	Presence of multiple modifications (sulfation, acetylation, methylation) and high diversity of linkages	Chemical synthesis is difficult and limited to small oligosaccharides due to the need of complex protecting and deprotecting strategies
	Site of attachment to protein/lipid	Requires glycan-protein and/or glycan-lipid characterization in addition to glycan structure
Glycan presentation and glycan-protein interactions	Presentation of an ensemble of different (often related) structures within a biological system or interaction	Studies must account for a population of glycans with similar structures, rather than an 'average' single structure
	Glycan-protein interactions often achieve high affinity and specificity by multivalency	Correct presentation of glycan and glycan-binding protein/domain(s) is critical for experimental design
	High torsional flexibility of glycans mediates presentation of a range of conformations for a particular glycan	Sequence of glycan is often not sufficient to characterize glycan-protein interactions; analysis of conformations and topologies should be considered

Table 1.1 **Challenges with decoding structure-function relationship of glycans**. Adapted from [32].

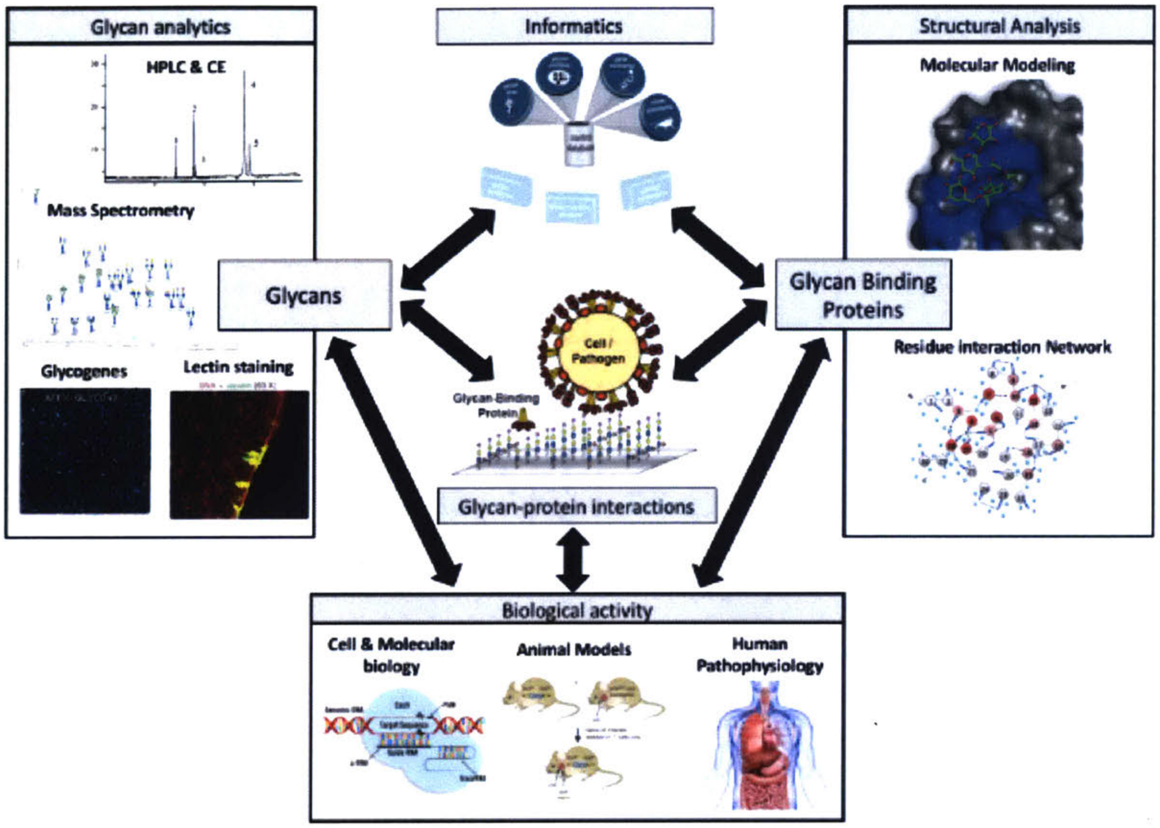


Figure 1.2 **Schematic depicting an integrated approach to decode structure function relationships of glycans.** Decoding the biological function of glycans requires an integration of multiple tools (representative examples of tools are depicted inside boxes). From the glycan perspective (left), attributes such as: glycan fine-structure or "sequence", the ensemble of glycans expressed (glycome) can be measured using analytical tools. From the GBP perspective (right), GBP-glycan interaction attributes such as specificity, and affinity as well as the structural basis for these interactions can be measured using glycan arrays and structural modeling. Biological activity should be assessed using experimental systems (i.e. cell or animal model systems) where glycan structure, GBPs, or glycan-protein can be perturbed (i.e. genetic, chemical, or enzymatic manipulation). Ultimately, integration and correlation across each of these axes are needed to converge on structure-function relationships.

Thesis outline

Motivated by these challenges discussed above, the goals of this thesis are to i) further develop and improve integrated approaches to study glycans and glycan-binding proteins, and ii) leverage these approaches to uncover new biological roles of glycans and GBPs in disease. Given that glycans exert their function through interactions with glycan binding proteins, decoding glycan structure-function relationships requires technologies to study glycans, glycan binding proteins and their interaction. Thus, this thesis is divided into two sections: Integrated approaches to study glycan binding proteins, and integrated approaches to study glycans. Furthermore, each section can be broadly divided into two parts: development of tools, and application of those tools.

In section one of this thesis work, I focus on integrated approaches to study glycan binding proteins (GBPs). Glycan binding proteins, through their direct interaction with glycans, decode the information encoded within the glycan milieu. Thus, understanding the structural basis for affinity and specificity of GBP-glycan interactions is crucial to decode structure-function relationships. To elucidate the structural basis for GBP-glycan interactions from the perspective of the glycan binding protein, it is important to characterize key structural determinants including (but not limited to): i) identification of the glycan binding site, ii) enumeration of key molecular contact made in the glycan-protein interface, and iii) identification of key functional residues facilitating the GBP-glycan interaction. To accomplish this, an approach that integrates structural analysis, with biochemical assays and functional readouts is required. Importantly, there is still a great need to develop new tools and improve approaches to characterize these structural determinants.

In part one, I develop tools and approaches to study the structural basis for GBP-glycan from the perspective of glycan binding proteins. Specifically, I implement a computational tool enabling inter-residue network analysis (alongside structural analyses, biochemical and functions assays) and investigate the tool's utility in identifying residues in the glycan binding site of a model GBP, FGF-2, that play a functional role in FGF-2's interaction with heparin (Chapter 2).

In part two, I apply this integrated approach to study influenza A virus (IAV) hemagglutinin (HA), a glycan binding protein that regulates host tropism, airborne transmissibility, and host immune recognition. Influenza pandemic pose a significant global threat. Influenza pandemics can occur when IAVs from non-human hosts, which are antigenically novel to humans, undergo a host switch and gain the ability to infect humans. In these cases, the HA has acquired mutations enabling a switch in its receptor specificity from avian glycan receptors to human glycan receptors. Here, using the integrated approach developed previously, I investigate the HA-glycan binding specificity of IAVs and uncover the structural determinants required for HA to switch its glycan receptor specificity (Chapter 3: H5N1, Chapter 4: H7N9). Furthermore, I integrate the previous two analyses with an informatics approach to assess antigenic novelty of H3 HAs (Chapter 5: H3N2, and Chapter 6: H3N8). Ultimately, the work here provides insights and strategies for influenza surveillance and enables early identification of IAVs with pandemic potential that are currently circulating in non-human hosts.

In section two of this thesis work, I turn my attention to the study of glycans and their biological function. Glycans are abundant on the cell surface, and at the cell-ECM interface where they mediate interactions between cells and their microenvironment. Importantly, glycans are expressed as a heterogeneous ensemble of structures where they interact with glycan binding proteins to mediate biological processes in an "analog" fashion. To elucidate the functional role of glycans, a key first step is to measure their structural attributes, including: glycan fine-structure or "sequence", the ensemble of glycans expressed (glycome), the glycoprotein to which it's attached, and its expression in cells or tissue. Due to glycan complexity there is no single tool capable of capturing all these features and multiple analytical tools and methods are needed. Importantly, measurements of glycan structure must be integrated with functional measurement from cell or organism model systems. Importantly, assessing glycan biological activity requires the ability to perturb the glycome and its interactions with GBPs (i.e. using genetic or chemical approaches). Ultimately, using this integrated approach (i.e. a functional glycomics approach) it is possible to converge on structure-function relationships. Owing to a growing interest in understanding cell-microenvironment interactions, there is still a great need to implement functional glycomics approaches, and apply them to decode the biological roles of glycans in disease processes.

In part one, I develop an integrated approach to characterize the cell surface milieu of glycans (Chapter 7). This approach integrates glycogene expression data, analytical tools such as mass spectrometry, and lectin arrays to characterize the glycome. I demonstrate the utility of a functional glycomics approach by applying it to characterize glycomics changes that cause metastatic progression in cell lines derived from a mouse model of multistage lung cancer.

In part two, I apply this functional glycomics framework to study the function that HSGAGs play in regulating cell-microenvironment interactions in a model of breast cancer stem cells. Almost all cancer related deaths are due to metastasis and resistance to therapy. The latter has been attributed to the existence of cancer stem cells, a cell population characterized by their increased resistance to chemotherapeutics and ability to seed new tumors and metastasis. Thus, targeting cancer stem cells therapeutically could lead to more durable clinical responses. Consequently, there is a great need to uncover the signaling pathways involved in cancer stem cell activity. The activity of cancer stems cells depends on their interactions with factors in their microenvironment. Despite knowledge that HSGAGs play a critical role in cell-microenvironment interactions, the function of HSGAGs in cancer stem cell activity is largely unknown. In Chapter 8, I study the role of HSGAGs in the activity of cancer stem cells. Specifically, I study how SULF1, an HSGAG modifying enzyme, regulates cell-microenvironment interactions to influence tumor initiation, metastasis, and maintenance of breast cancer stem cells.

Next, I leverage my expertise in integrated analytics to explore an intellectually adjacent area, namely, developing an analytical framework to assess product consistency/potency for therapeutic extracts. Exposure to certain environments with high microbial exposures, such as traditional farm environments, can protect against allergic type disease. Recent evidence suggests that the protective effect can be conferred through exposure to dust from cow sheds. This observation has led to a burgeoning interest in using dust extracts as therapeutics to prevent allergic-type disease. As dust extracts are complex mixtures comprised of various bioactive molecules, including complex polysaccharides, they pose a significant regulatory challenge. Here, I conceptualize an integrated analytical strategy to assess product consistency/potency for therapeutic farm dust extracts, thus addressing regulatory concerns surrounding the development of farm dust extract-derived therapeutics (Chapter 9).

The remainder of this chapter will focus on detailing the motivation and provide the relevant introductory material for the applications areas discussed in this thesis:

1. Integrated approaches to study hemagglutinin and hemagglutinin-glycan interactions: A strategy to improve pandemic risk assessments of IAVs from non-human hosts.
2. Integrated approaches to study cell surface glycans: Uncovering the role of heparan sulfate glycosaminoglycans in modulating cancer stem cell activity through regulation of cell-microenvironment interactions

Section 1: Integrated approaches to study hemagglutinin and hemagglutinin-glycan interactions: A strategy to improve pandemic risk assessments of influenza from non-human hosts

Motivation and Objectives

Influenza A Virus (IAV) poses a significant public health threat. Each year influenza epidemics result in approximately 3-5 million cases of illness and between 200,000 and 500,000 deaths worldwide. In high-risk groups, such as elderly, young, pregnant or immunocompromised individuals, influenza infection can result in hospitalization or death. Perhaps more concerning than annual epidemics are influenza pandemics. There have been four well-documented influenza pandemics over the last century; the 1918 H1N1 Spanish flu, which killed ~20-40 million people, the 1957 H2N2 Asian flu, which killed 2-3 million, the 1968 H3N2 Hong Kong flu, which killed ~2 million, and the 2009 swine flu pandemic, which killed ~200,000 people [33,34]. Because pandemics pose such as serious public health threat, there is significant motivation to develop tools and strategies to predict the emergence of future influenza pandemics. Successfully doing so could enable early interventions, such as the creation of vaccines or therapeutics, which could limit the devastating morbidity and mortality associated with pandemics. In this context, it is important to understand the molecular determinants of IAV pathogenesis.

Pandemics are the result of antigenic shift, where an antigenically novel IAV strain is introduced into the human population [35]. Because birds are the natural reservoir for influenza, pandemics are thought to emerge when avian influenza viruses (or avian-derived viral components via viral reassortment) gain the ability to infect antigenically naïve human populations. Fortunately, this event is rare as IAVs that infect birds cannot infect humans without acquiring mutations.

IAV host tropism and airborne transmissibility in humans is regulated by the viral coat protein hemagglutinin (HA), which regulates the first step of infection through its interaction with sialylated glycan receptors on the cell surface [36]. Importantly, the glycan binding specificity of HA is a critical determinant of host tropism. Humans express $\alpha 2 \rightarrow 6$-linked sialylated glycans and birds express $\alpha 2 \rightarrow 3$-linked glycans. In order for avian-adapted IAVs to infect humans, mutations must occur in HA that facilitate a shift in glycan binding specificity from $\alpha 2 \rightarrow 3$ to $\alpha 2 \rightarrow 6$-linked glycans [37].

Using sequence information from circulating IAVs, a molecular surveillance approach enables our ability to predict the emergence of new viral pandemics by identifying IAV strains circulating in non-human hosts that pose a significant pandemic threat [38,39]. In the context of HA, an effective molecular surveillance approach depends on uncovering the structural determinants that underlie HA-glycan specificity. Accomplishing this allows for the identification of mutations that could give rise to a switch in receptor specificity.

Much of the current research studying the structural determinants of HA-glycan interactions has significant limitations. Many studies in this field adhere to an over-simplified view of the HA-glycan interaction where specificity is governed only by the terminal sialic acid linkage. Furthermore, many current experimental tools to study HA-glycan interaction do not account for important features such as glycan conformation, multivalent presentation of the HA, or HA-glycan affinity [15,37,40]. By ignoring these properties of HA-glycan interactions, it has been difficult to define the structural determinants of HA-glycan interactions and correlate these with outcomes like host tropism switch or airborne transmissibility.

Previous work performed in our lab uncovered that HA-glycan specificity depends on the topological presentation of the glycan, where avian-adapted HAs recognize $\alpha 2 \rightarrow 3$ linked sialylated glycans that adopt a cone-like topology and human-adapted HAs recognize $\alpha 2 \rightarrow 6$ linked glycan that adopt an umbrella-like topology [15]. Furthermore, our lab has previously developed biochemical assays capable of assessing the quantitative affinity of multivalent HAs to bind to avian or human glycan receptors that adopt their respective topologies [41].

Leveraging these insights, the work in Chapters 3-6 of this thesis focuses on studying Hemagglutinin and the HA-glycan interaction to uncover insights and develop new strategies to aid in molecular surveillance of influenza. I apply a unique approach, which integrates structural modeling of the HA-glycan receptor complex, inter-residue network analysis of HA, biochemical analysis of HA-glycan interactions, and informatics analysis of antigenic properties. Specifically, the following are the objectives of Section 1:

a. Identify the structural determinants for naturally evolving H5N1 HA to switch its receptor specificity from avian to human (Chapter 3)
b. Determine the physiologic glycan receptor binding properties of a 2013 outbreak H7N9 HA and identify the mutations required to enable binding to human glycan receptors. (Chapter 4)
c. Develop bioinformatics tools and experimental methods capable of measuring antigenic properties of HA. Leverage these tools to identify avian and swine-adapted H3s that could re-emerge into the human population and potentially cause a pandemic (Chapter 5).
d. Characterize the glycan binding specificity of an HA isolated from the 2011 New England harbor seal H3N8 (Chapter 6)

Influenza virus structure

Influenza is a single stranded negative-sense RNA virus that belongs to the family *Orthomyxoviridae*. Influenza virions are spherical or elongated particles that are approximately 80-120nm in diameter. Although there are three influenza genera (Influenza A, B, C), Influenza A viruses will be the focus of this thesis as they are predominantly responsible for seasonal outbreaks, and pandemics. The influenza A genome is comprised of eight segments encoding 10 proteins (Figure 1.3) which are discussed below in detail. Unless cited otherwise, information for this section was obtained from [40,42] :

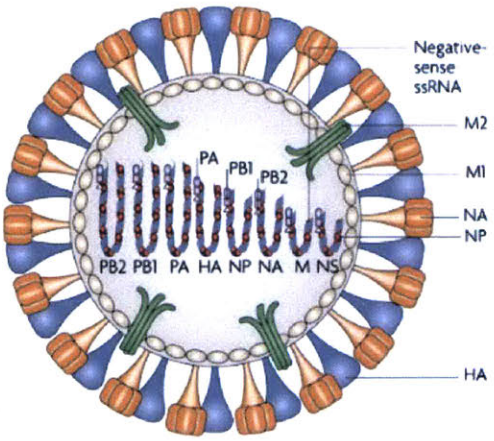


Figure 1.3 **Influenza A Virus structure**. Influenza is a single stranded negative sense RNA virus which is composed of 8 gene segments. The function of each viral component is enumerated in the text. Adapted from [42]

Viral glycoproteins

Two viral glycoproteins coat the surface of influenza virions: Hemagglutinin (HA) and Neuraminidase (NA). Importantly, both of these envelope glycoproteins are the major site of immune recognition. In fact, Influenza A Virus subtypes are classified serologically based on the presence of HA and NA subtypes (e.g. H1N1, H3N2, or H5N1). There are 16 different HAs and 9 different NAs, yielding 144 possible combinations. Only a subset of HA/NA combination have been observed in nature. While all these subtypes circulate in waterfowl, only a few have been observed in the human population [36]. H1N1, H2N2, and H3N2 have been observed to infect humans, and initially gained foothold in human populations during past pandemic outbreaks: 1918 H1N1, 1957 H2N2, 1968 H3N2 [43]. Importantly, H1N1 and H3N2 continue to circulate in humans, causing seasonal outbreaks. Recently, several cases of highly pathogenic avian influenza viruses, like H5N1, H7N7 and H9N2 have been shown to infect humans [44–46]. These outbreaks are usually small, and result from direct contact with infected animals; however, their high case fatality rate (e.g. 60% for H5N1) make these transmission events worrisome from a public health standpoint [46].

Functionally, HA and NA coat the viral envelope, giving rise to roughly 500 spike-like projections coming off of the capsid. The HA spike is composed of an HA trimer, while the NA spikes are tetramers. The HA is responsible for entry into the host cell, with its two key functions being viral attachment to host glycan receptors on the cell surface, and membrane fusion in the endosome. The NA protein cleaves cell surface sialic acids. NA allows for the release of newly formed virions, which also prevents the re-infection of the virus producing cell [47].

Internal proteins

The influenza A virus contains seven internal proteins: RNA polymerase (PB1, PB2, PA), Nucleoprotein (NP), Matrix protein (M1 and M2), and Non-structural proteins (NS1 and NS2/NEP). This thesis predominately focuses on the HA viral glycoprotein, so the function of each internal protein will only be discussed in brief [36,40].

Viral Ribonucleoprotein (RNP) complex

Influenza's eight gene segments of single stranded RNA are packaged as ribonucleoprotein complexes (RNPs). Each RNP is comprised of a viral polymerase complex (composed of PB1, PB2, and PA) bound to a short hairpin of vRNA, and multiple copies of viral nucleoprotein (NP) that also bind vRNA, acting as a structural protein [48].

Non-structural proteins

Non-structural protein, NS1, is known to play several important roles such as enhancement of viral mRNA translation, inhibition of host mRNA processing, and inhibition of the host cell immune response. NS2, which is also known as nuclear export protein (NEP), facilitates nuclear export of viral RNP complexes.

M proteins:

The M1 matrix protein is involved in the export of viral RNP from the nucleus to the cytosol. M2 is an ion channel responsible for the acidification of the endosome, which is critical for fusion of the viral and host membranes, and release of vRNPs in the cytosol.

Influenza viral life cycle

The influenza lifecycle, shown in Figure 1.4, begins with a virion associating with the host cell surface, where the viral hemagglutinin binds to sialylated glycan receptors [36,40]. Next, the bound virus is endocytosed, predominantly through clatharin-coated pits, but other mechanisms have been described [49]. Next, the endosome is acidified by the M2 ion channel matrix protein. This promotes a conformational change in the HA protein which mediates fusion of the viral and endosomal membranes. Following endosomal fusion, the uncoated viral ribonucleoprotein (RNPs) complex is released into the cytosol of the host cell. The ribonucleoprotein complex is transported into the nucleus, where incoming negative sense viral RNA (vRNA) is transcribed into mRNA. During viral replication, a full-length complementary RNA is first made (i.e. a positive sense copy of the vRNA) then used as a template to produce more vRNA. Next, the newly synthesized viral RNAs are exported to the cytoplasm by NEP, where the viral proteins are translated. The viral proteins are expressed, processed and targeted to budding sites on the host cell membrane. The viral surface glycoproteins, protein complexes, RNPs, vRNAs are assembled in viral particles and bud from the host cell membrane. The viral neuraminidase cleaves cell surface sialic acid receptors on host cells, which facilitate the release of new virions and simultaneously prevents re-infection of the same cell [36].

The work in this thesis focuses on the first step in the viral lifecycle; the binding of viral HA to host cell surface glycan receptors. Specifically, this work focuses on the viral hemagglutinin-glycan receptor interaction and its role in tropism and transmissibility in human hosts.

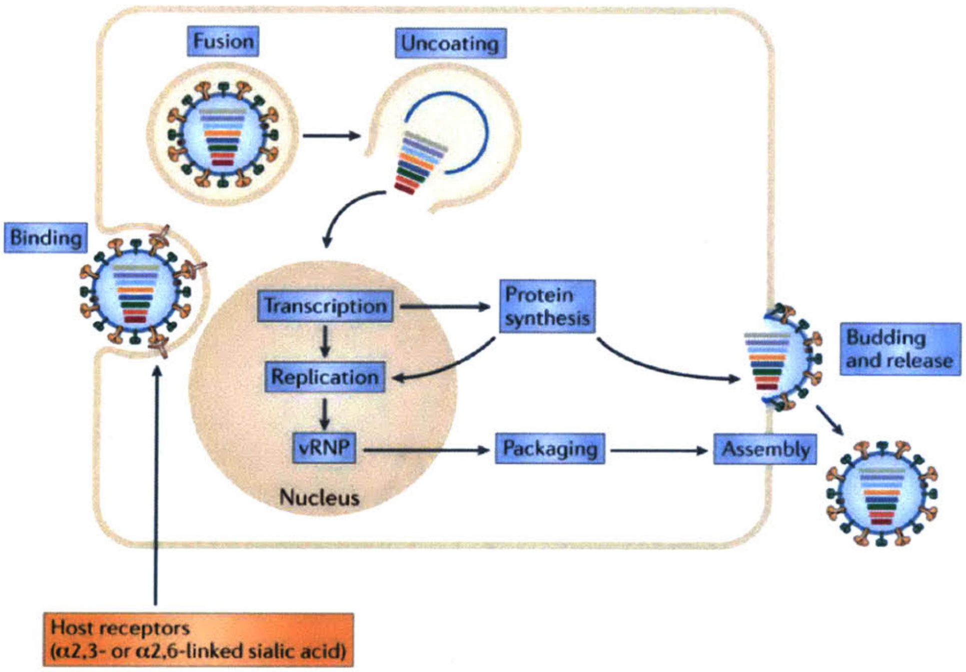


Figure 1.4 **Pathway to Viral Infection by Influenza A.** Viral infection is initiated by binding of viral glycoproteins to the sialylated receptors on the host cell membranes. Entry of the virus into the host is facilitated by endocytosis. Next the viral capsid fuses with the host endosome, releasing its genetic content into the host. Viral RNA enters the host nucleus where transcription and replication occur via RNA polymerase. Viral mRNA is produced and used to synthesize viral proteins in the cytoplasm which are then assembled into viral ribonucleoproteins (vRNPs) in the nucleus. Virus particles are assembled at the cell membrane where budding occurs to release the particles into the extracellular space. Adapted from [36].

Influenza virus ecology

Influenza A viral ecology is complex, yet critical for understanding how pandemics emerge. Influenza is primarily a zoonotic pathogen and has been shown to infect a variety of animals including birds, pigs, dogs, horses and humans. In its natural ecology, the Influenza A virus exists as a commensal in the gut of migratory aquatic birds, including shorebirds and water fowl [36]. In these hosts, the virus does not cause disease and thus a significant amount of genetic diversity is introduced. Mutations can arise that enable the influenza viruses to infect domestic mammals or humans (Figure 1.5). In order to "jump" into domestic mammals or humans, aquatic birds must first contact and infect domestic birds, such as chickens or ducks. Subsequently, these avian viruses can be transmitted from domestic fowl to domestic pigs, which are susceptible to infection from both avian and human viruses. Swine are hypothesized to act as intermediate hosts that facilitate

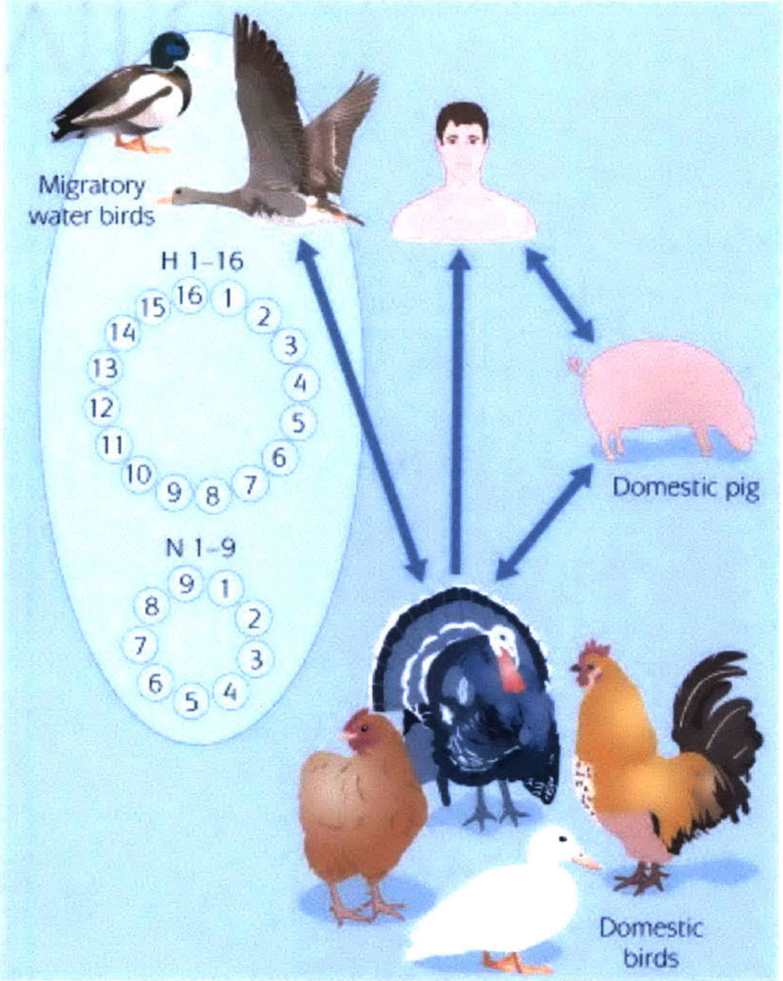


Figure 1.5 **Origin of Antigenic Shift and Pandemic Influenza.** The majority of influenza A viruses and existing HA subtypes (H1-16 and N1-9) circulate in migratory water birds. Two ways a virus with a new HA subtype may be introduced into the human population are direct transmission of an avian virus to humans or genetic re-assortment between an avian and human virus through infection of domestic pigs. Adapted from [297]

viral re-assortment by acting as genetic "mixing vessels", leading to flu novel subtypes that can infect humans [50,51].

Occasionally, avian or swine influenza viruses gain the ability to infect human hosts (Figure 1.5). This can result in a pandemic in the instance that the new viral subtype is antigenically novel to humans (known as antigenic shift) [34]. In fact, it is thought that the previous pandemics all originated from avian or swine viruses, which were antigenically novel to humans, that gained the ability to infect humans [36].

Origin of influenza pandemics

There are three hypothesized mechanisms by which new viral subtypes emerge in human populations and give rise to a pandemic:

1. **Direct transmission from avian to human hosts**. In this instance, viruses acquire mutations in key viral proteins such as hemagglutinin (enabling human infection) or viral polymerases (enabling replication in human tissue) which are necessary for human adaptation.
2. **Genetic reassortment** can occur when two Influenza A viruses infect the same host cell. Upon infection, both vRNA sets are replicated and during viral assembly the RNA segments from the two viral strains become mixed to produce a novel virus with a unique combination of genes [34]. Pigs are often considered "mixing vessels" for viral reassortment as they have the capability to be infected with both avian and human viruses [51]. The 2009 swine flu pandemic H1N1, as thought to have emerged through genetic reassortment of bird, swine and human viruses [52].
3. **Reintroduction of an "old" strain into the human population**. Viral strains, for a variety of reasons, may disappear from the human population, and can possibly re-emerge. If this occurs when the immunity in the infected population has waned, this virus will be perceived as antigenically novel. This was thought to be the case for the 1977 H1N1 "Russian flu", which was found to be genetically identical to an H1N1 that circulated in humans in the 1950s [53].

Although not discussed in detail here, antigenic novelty can yield pandemics and contribute to seasonal epidemics. Antigenic novelty arises through two major mechanisms; Antigenic shift (discussed above), and antigenic drift [54]. Antigenic drift refers to the gradual accumulation of mutations on the antigenic domains of Hemagglutinin or Neuraminidase. The replication of vRNA is an error prone process. In fact, the mutation rate for influenza is ~1/100,000 nucleotides, which means many of the nascent vRNA copies will likely contain one or more mutations [55]. Pressure from host immune systems can drive the gradual evolution of antigenic sites on viruses. In cases of significant antigenic divergence, it is possible that a population may not be protected by their pre-existing immunity. This can yield more severe mortality and morbidity events in a given flu season [56]. The rapid evolution of influenza viruses via shift or drift means constant monitoring of viral evolution is required (i.e. surveillance). Surveillance is important and serves to inform seasonal vaccine composition and helps monitor the geographic spread of influenza. It can also aid in predicting supply needs and can help prioritize allocation of vaccines.

Host glycan receptors

Glycans are known to play a critical role in host-pathogen interactions. One of the most well-characterized interactions is between Influenza A Virus (IAV) hemagglutinin and the sialylated glycans receptors on the surface of cells. As mentioned previously, the first critical step of viral entry is the binding of HA to glycan receptors. Importantly, the glycan-receptor specificity of IAV HA determines host adaptation and tissue tropism.

IAVs circulate as commensals in the gut of aquatic birds. In avian hosts, the gut and respiratory tract predominantly express N- and O- linked glycans which contain terminal sialic acids that are $\alpha2\rightarrow3$-linked to galactose (referred to interchangeably as avian receptors) (Figure 1.6) [57]. In humans, however, the primary site of infection is the upper respiratory tract, which predominately express N- and O-linked sialylated glycans with an $\alpha2\rightarrow6$ terminal linkage (referred to interchangeably as human receptors) [15]. Importantly, hemagglutinins derived from avian-adapted viruses have specificity for α2→3-linked sialylated glycans, while HAs from human-adapted viruses have specificity for α2→6-linked sialylated glycans.

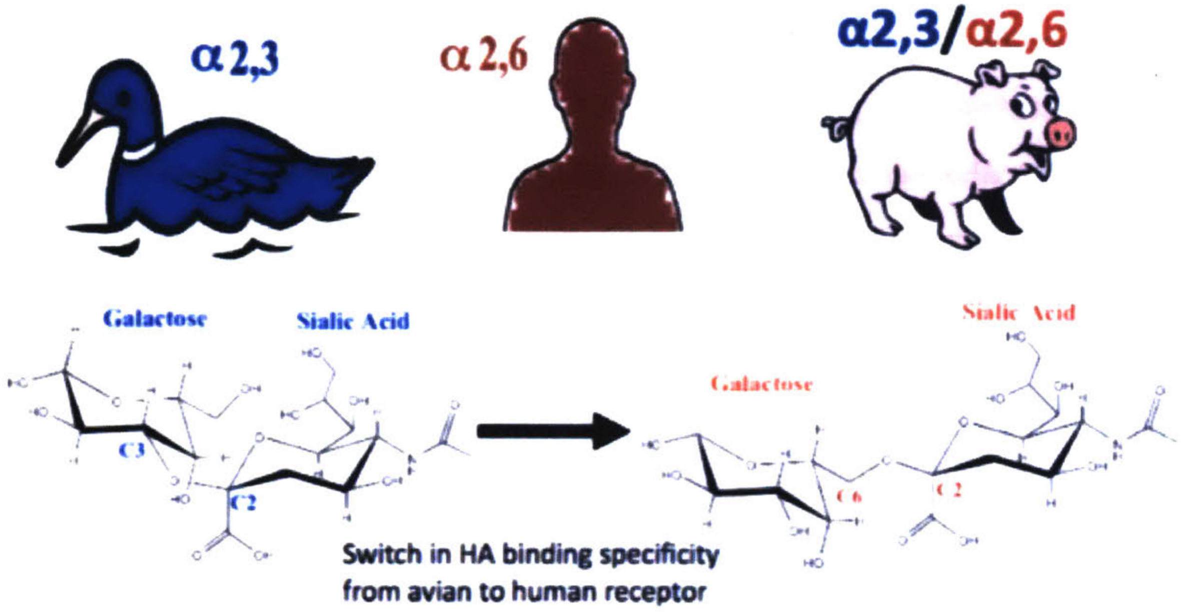


Figure 1.6 **Glycan Receptors for Influenza HA in Various Hosts**. (Bottom) Chemical structure of the glycan receptors. Sialic acid is linked to galactose via α2-3 or α2-6 linkage. α2-3 and α2-6-linked glycan are the avian and human receptors for HA, respectively. Viral Host tropism is determined by the glycan receptor distribution in the host. (Top) Birds express α2-3, Human express α2-6, and pigs express α2-3/α2-6. A switch in HAs receptor binding specificity from α2-3 to α2-6 is necessary for human adaptation of the virus. Adapted from [40].

The distribution of glycan receptors in the upper respiratory tract of human tissue is also important, especially when considering cellular tropism. In order to investigate this question, lectins (a class of glycan binding proteins with well-characterized specificity) have been used to investigate the physiologic distribution of glycans in human lung tissue. Using two lectins, SNA-I (specificity for sialic acid that is α2→6-linked to Gal or GalNAc) and MAL-II (specificity for α2→3-linked glycans), the upper respiratory tract of human, namely the tracheal epithelium, was shown to express α2→6-linked glycans [15]. Conversely, α2→3-linked sialylated glycans were mostly present in alveolus in the lower respiratory tract. Indeed this distribution of α2→3 vs α2→6 glycan receptors strongly correlated with cellular tropism of human-adapted vs avian-adapted viruses [58]. HAs that are derived from human-adapted viruses bind to non-ciliated goblet cells and ciliated epithelial cells in the upper respiratory tract, whereas HAs from avian-adapted viruses do not bind the upper respiratory tract and localize to deep lung and alveolar tissue. While the exact mechanism is unknown, the cellular tropism of pandemic viruses (i.e. binding to non-ciliated goblet cells and ciliated epithelial cells) seems to be important for their efficient airborne transmission [58].

In contrast to bird and humans, the epithelial cells on the upper respiratory tracts of pigs express both $\alpha 2 \rightarrow 3$ and $\alpha 2 \rightarrow 6$-linked sialylated glycan receptors. Consequently, pigs can be infected with both avian- and human-adapted IAV. For this reason, pigs are considered an evolutionarily intermediate host enabling viruses to jump from domestic birds to humans via acquisition of mutations or by acting as a “mixing vessel” for genetic re-assortment of IAV (Figure 1.6).

Influenza A virus hemagglutinin: Structure and interaction with glycans

HA is the most abundant protein of the influenza viral surface. It plays several critical roles during the viral lifecycle and disease pathogenesis:

1. HA mediates viral entry and endosomal fusion.
2. HA determines host tropism through it preference for $\alpha 2 \rightarrow 3$ or $\alpha 2 \rightarrow 6$ linked sialylated glycan receptors.
3. HA-glycan specificity regulates respiratory droplet transmission.
4. HA is the major site of host immune recognition.

HA is produced as a precursor polypeptide (HA0), which is a type 1 transmembrane glycoprotein ~ 550 amino acids. The HA0 precursor becomes activated via proteolytic cleavage into two peptides HA1 and HA2, which are disulfide linked. Ultimately, HA is displayed on the viral surface as a trimer of three monomers of HA1-HA2. The first crystal structure of the HA ectodomain was reported in the early 1980s [59]. The protein is generally divided into two domains: The membrane proximal stem domain, and the membrane distal globular head domain (**Error! Reference source not found.**a). The globular head is of interest in this thesis as it contains the glycan receptor binding site (RBS) and the antigenic region [43].

Molecular insights into the HA-glycan interaction

The glycan receptor binding site (RBS) is a shallow pocket present on the tip of each monomer in the globular head region. The RBS is composed of a base, including several conserved amino acids (Tyr-98, Trp-153, His-183, and Tyr-195), and three structural elements around the edges of the pocket (the 130- and 220- loop, and the 190-α helix) (**Error! Reference source not found.**a) [36]. When the sialylated glycan receptor is bound to HA, the terminal sialic acid residue makes critical hydrogen bonding contacts with the base residue, Y98, and the 130 loop. The monosaccharides distal from the terminal sialic acid interact with the 220-loop and the 190 helix. Furthermore, W153, H183 and Y195 make additional van der Waals interactions with the glycan receptor (**Error! Reference source not found.**a).

HAs specificity for α2→3 versus α2→6 sialylated glycans determines hosts tropism. Therefore, it is important to understand the structural determinants which govern the receptor specificity. Crystal structures have been obtained for avian-adapted and human-adapted viruses bound to LSTa (an α2→3 linked pentasaccharide) or LSTc (an α2→6 linked pentasaccharide)(**Error! Reference source not found.**). When bound to HA the avian and human receptors adopt different conformations. Co-crystal structures of avian-adapted HAs in complex with LSTa, shows that the α2→3 linked glycans exhibit a *trans* conformation, where the Gal and GlcNAc residues are in an extended conformation relative to the sialic acid residue. Furthermore, the glycosidic oxygen atom faces the 220-loop, which presents hydrogen bond between the Gln 226 and the 4-hydroxy group of Gal-2 and the oxygen in the glycosidic linkage (Figure 1.7b) [36]. These interactions are present in all avian-adapted HAs. In contrast, the α2→6 human glycan receptor (LSTc) bound to human-adapted HAs show a *cis* conformation and appear "folded". In this conformation the glycosidic oxygen points away from the

RBS and the hydrophobic C6 atom of galactose points towards the RBS base (Figure 1.7c).

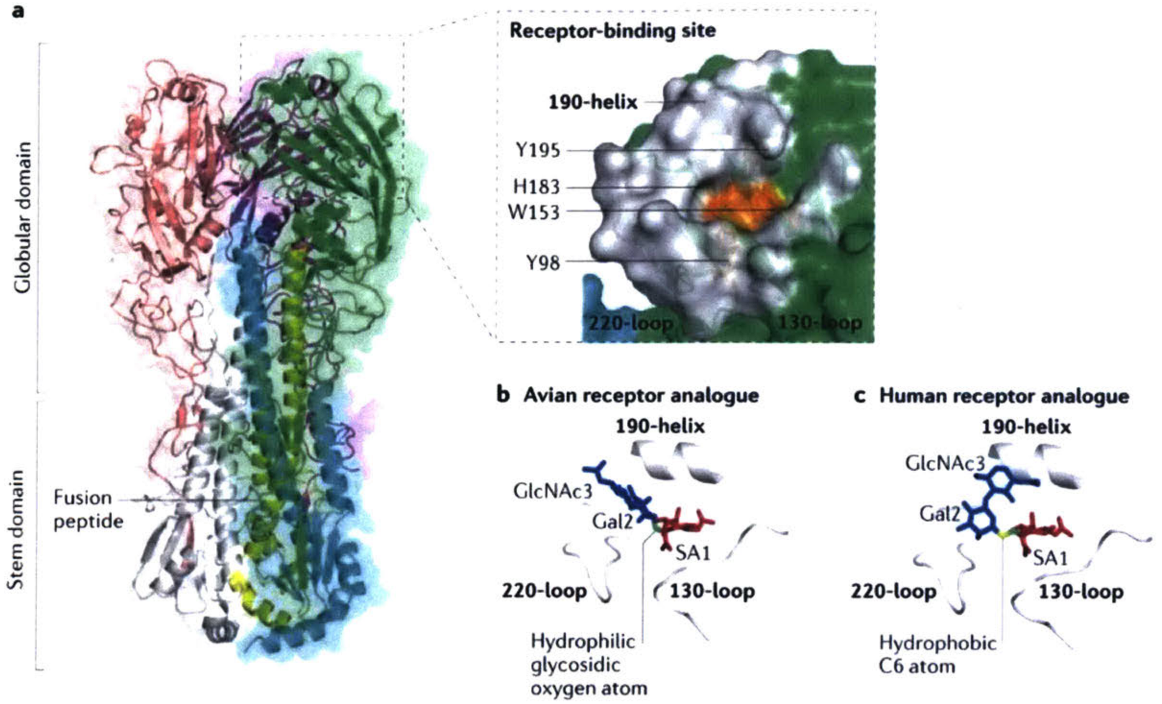


Figure 1.7 **Hemagglutinin and the Glycan Receptor binding site. a)** Shown here is the crystal structure of HA (PDB ID= 4JUG). The two domains of the HA ectodomain are labelled: the globular head domain and the stem domain. The globular head domain contains the receptor-binding site which forms a shallow pocket comprising three structural elements: the 130-loop, 190-helix and 220-loop. Four highly conserved residues (Y98, W153, H183 and Y195) form the base of the receptor-binding site (shown in orange). b,c) Human and avian receptor analogues show differing structural conformation upon binding to HA (PDB ID= 4JUH and 4JUJ). The three terminal monosaccharide of the glycan receptor are shown: terminal sialic acid SA1, galactose Gal2, and N-acetylglucosamine GlcNAc3. [36]

As mentioned earlier, pandemics often emerge from avian viruses that undergo a host switch. In this case, avian-adapted HAs must switch their glycan receptor binding preferences from α2→3 to α2→6-linked glycans. For this to occur, influenza viruses must acquire mutations, which cause changes in the HA glycan receptor binding sites that facilitate a switch in receptor binding specificity. Therefore, understanding the

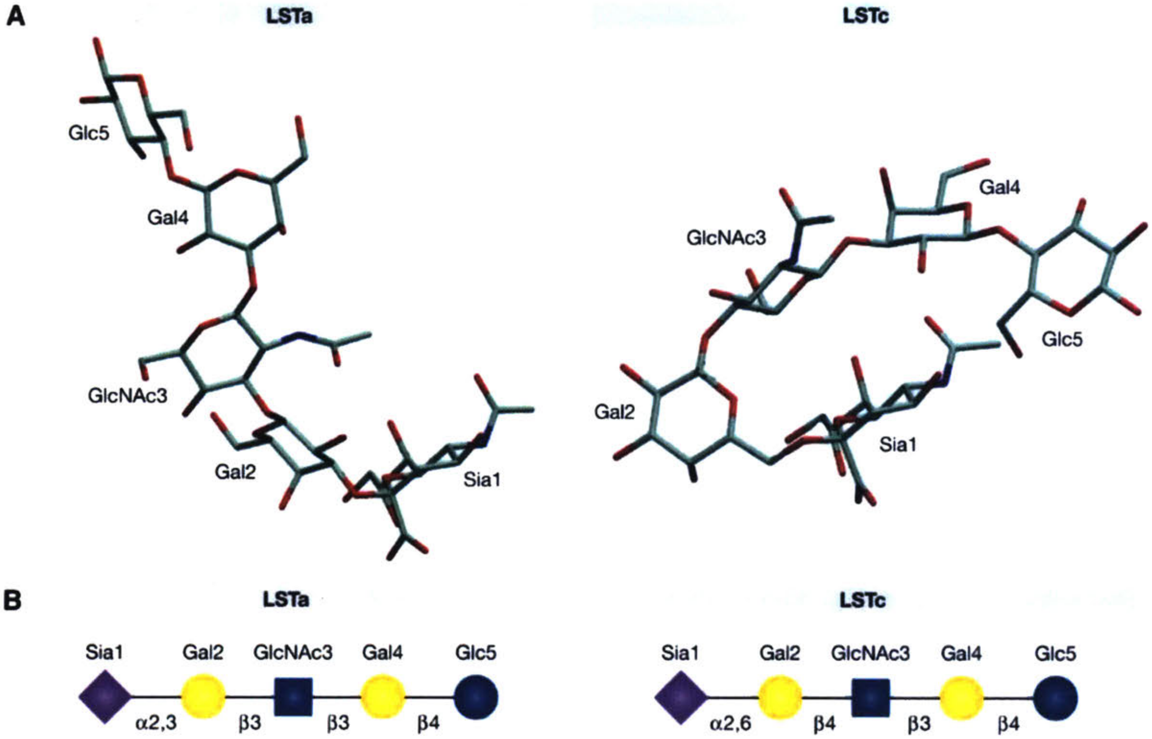
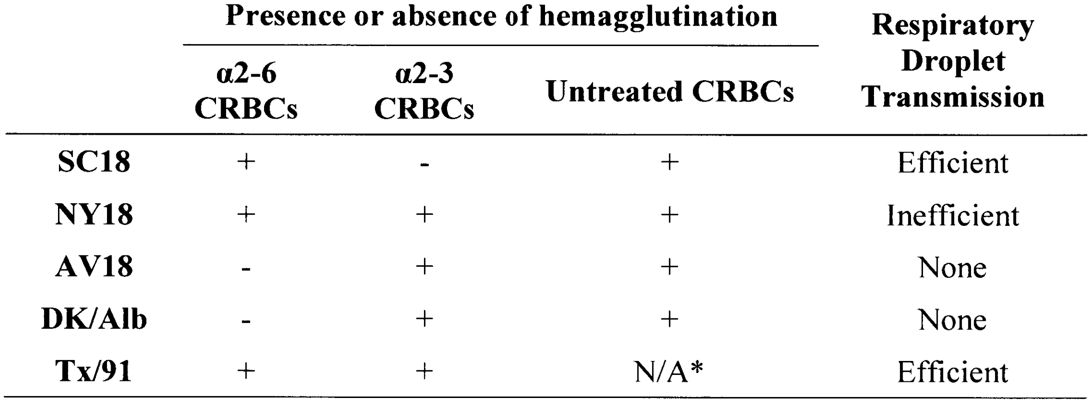


Figure 1.8 **Human and Avian Receptor Analogs in Free Confirmation.** a) Free confirmation of pentasaccharides LSTa, the avian receptor analog, and LSTc , the human receptor analog, show LSTa in *trans* conformation and LSTc in *cis* conformation. b) Cartoon representations of LSTa and LSTc structures are show (CFG nomenclature). Adapted from [57]

structural determinants and specific mutations that govern a “switch” are critical to enumerate as they enable identification of viruses with ‘pandemic potential’.

Challenges in elucidating HA-glycan interactions

Studying the “host switch” and human-to-human transmission is critical for our understanding of influenza pathogenesis and viral pandemics. Towards this, much work has been conducted using reverse genetics approaches to study these properties of Influenza A biology. Using reverse genetics, it is possible to recreate viruses and test their infection, replication, and transmission properties in ferrets [60,61]. Ferrets contain similar glycan receptors as humans and exhibit similar disease symptoms. In ferrets, various modes of transmission such as contact or respiratory droplet transmission can be studied by controlling ferret contact (i.e. co-housing or housed separately with a perforated wall, respectively) [40]. Using this model system, previous studies have been performed to investigate the virulence & transmissibility properties of single gene reassortments of the 1918 pandemic H1N1 (A/South Carolina/1/18 or SC18). Initially, studies demonstrated that the HA protein played a dominant role in IAV virulence (relative to other genes like NA or PB2) [62,63]. In order to further enumerate the role of HA in virulence, viruses were created that contained identical genes but swapped in various naturally occurring or mutant HAs. These studies investigated respiratory droplet transmission and viral replication in ferrets, and glycan receptor binding properties were tested using an RBC hemagglutination assay. Three HAs were tested, HA derived from SC18, NY18 (a single point mutant of SC18), and AV18 (a double point mutant) [37]. Using RBC agglutination, it was determined SC18 binds α2→6 glycans, NY18 binds both α2→3/α2→6, and AV18 only binds α2→3 glycans. SC18 was able to transmit efficiently in respiratory droplets, while NY18 was inefficient, and AV18 did not transmit (Figure 1.9).



	Presence or absence of hemagglutination			Respiratory Droplet Transmission
	α2-6 CRBCs	α2-3 CRBCs	Untreated CRBCs	
SC18	+	-	+	Efficient
NY18	+	+	+	Inefficient
AV18	-	+	+	None
DK/Alb	-	+	+	None
Tx/91	+	+	N/A*	Efficient

Figure 1.9 **Glycan receptor binding specificity and respiratory droplet transmission of H1N1 Viral HA reassortants.** Note that NY18 is a single amino acid mutant of SC18 (D225G) and AV18 is a double amino acid mutant of SC18 (E190D/D225G). Adapted from [37].

These results suggested that the loss of α2→3 binding is necessary for efficient transmission, while gaining specificity α2→6 was necessary but not sufficient to enable efficient transmission. Interestingly, confounding cases emerged where mixed α2→3/α2→6 binders showed variable rates of transmission. For instance, NY18 was a mixed binder and showed inefficient transmission, whereas an HA from A/Texas/36/91 (or TX91) was a mixed binder but transmitted efficiently [37]. These confounding results led many to hypothesize that there was something more complex governing transmissibility beyond this simple α2→3/α2→6 paradigm.

Integrated approach to decode HA-glycan receptor specificity

As discussed above, aspects of glycan chemical properties and their fundamental mode of necessitate a unique approach to determine biological function. Glycan-protein interactions are low affinity (uM-mM), and glycans possess a high degree of structural complexity. Because of this complexity, data across multiple lines of inquiry (structural, biochemical, and functional) should be integrated to elucidate structure-function relationships of glycans. Previously in our lab, integrated approaches were used in order to better understand avian- and human-adapted HA glycan binding specificity [15]. This approach used complementary methodologies to decipher fine structural properties of sialylated glycan receptors and bridge structural analysis with biochemical analysis of HA-glycan interactions (Figure 1.10).

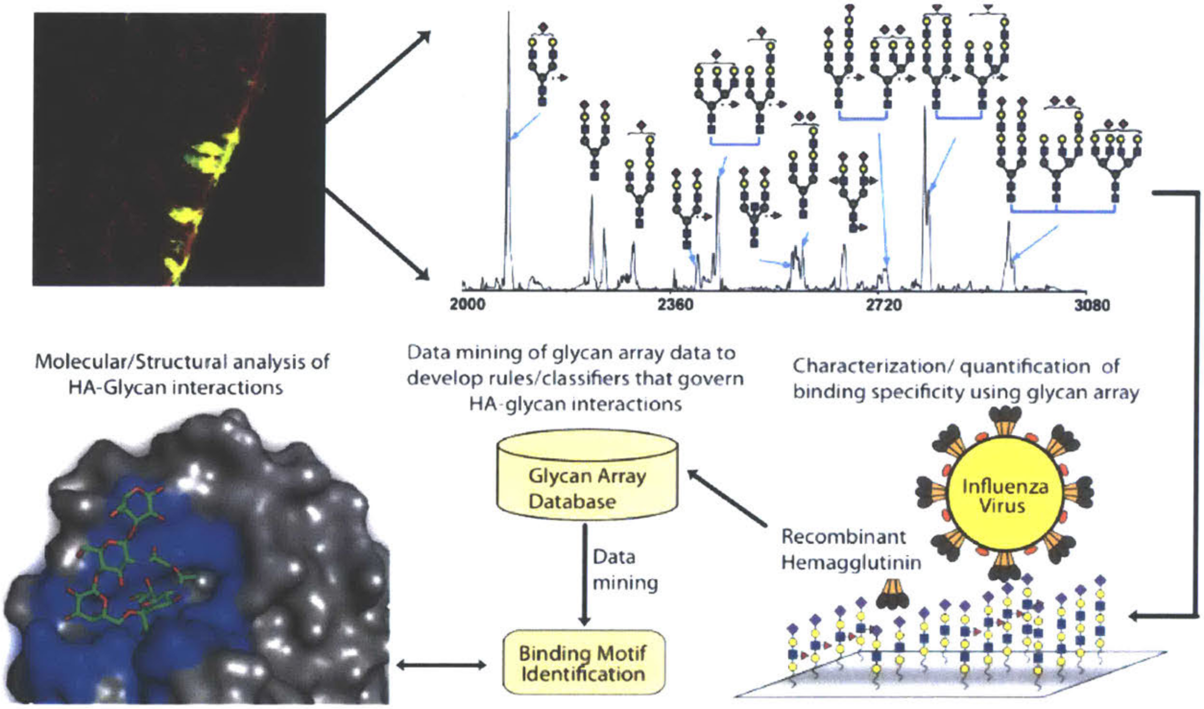
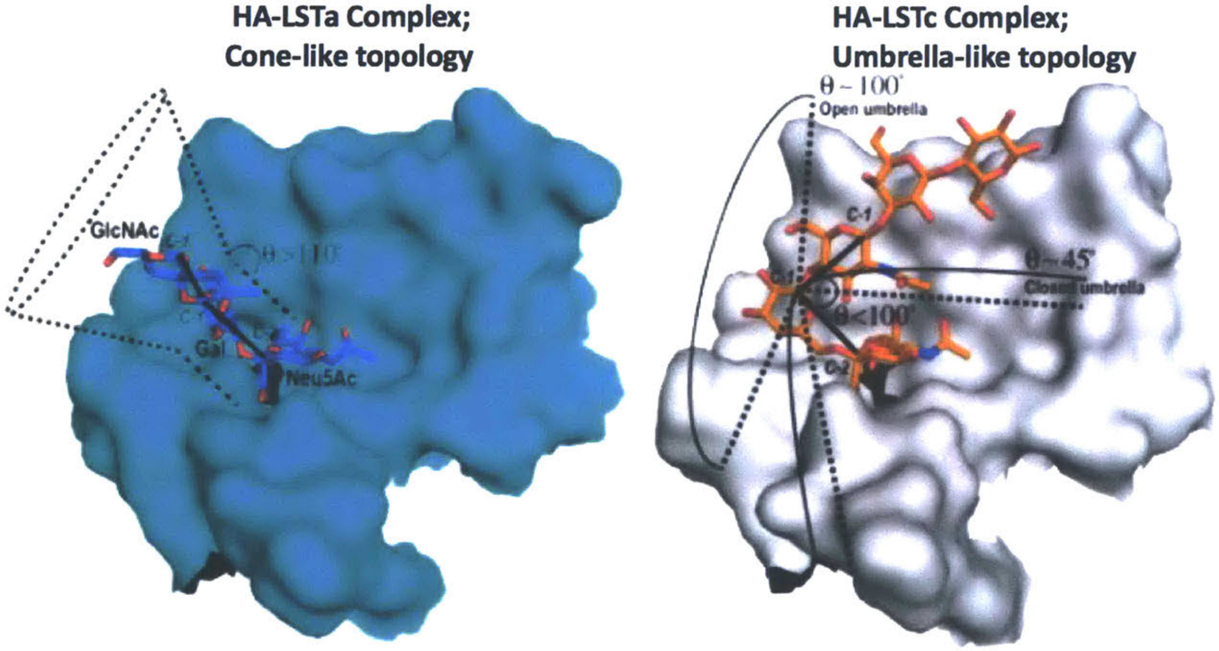

2000
2360
2720
3080
Molecular/Structural analysis of HA-Glycan interactions
Data mining of glycan array data to develop rules/classifiers that govern HA-glycan interactions
Characterization/ quantification of binding specificity using glycan array
Glycan Array Database
Data mining
Binding Motif Identification
Influenza Virus
Recombinant Hemagglutinin

Figure 1.9 **Integrated analyses to elucidate the key determinants of HA-glycan interactions that govern Influenza A pathogenesis.** (T,L) Lectin staining of the human upper respiratory tissues was used to determine the tissue level distribution of α2-3, and α2-6 sialylated N- and O- linked glycans. (T,R) Detailed structural information on the composition of the glycan pool isolated from human bronchial epithelial cells was determined using MALDI-MS in combination with sialidases. (B,R) Glycan binding specificity was determined using recombinant HA or whole virus on a glycan array platform. The array platform was designed to incorporate target structures based on their predominant expression in the upper respiratory tract. (B,M) The HA/Virus binding data were mined to obtain rules or classifiers that govern the glycan binding specificity of human adapted or wild type HAs. (B,L) The rules obtained were corroborated using X-ray co- crystal structures of HA-glycan complexes and molecular simulation of HA-glycan interactions. (T=Top, B=Bottom, M= Middle, R=Right, L=Left) Adapted from [76].

Briefly, the upper respiratory tract is the primary target for human-to-human transmission of influenza A virus, thus characterizing the diversity and distribution of the sialylated glycan receptors in this tissue was critical to define the viruses target glycan receptor. First, the fine structure of N-linked glycans isolated from a human bronchial epithelial (HBE) cell line was determined and quantified using MALDI-MS/MS coupled with sialidase enzyme treatments. Furthermore, a panel of lectins was used to stain tissue sections of upper respiratory tract to determine the physiological distribution of glycan receptors in tissues that human-adapted HAs infect. Using these complementary tools, it was found that both HBEs and tracheal tissue showed an abundance of multiantennary sialylated, predominantly α2→6 linked, glycans with long branches composed of multiple lactosamine repeats [15]. Next, a glycan array platform was used to determine the identity of the structural motifs necessary for human adapted HA binding. The glycan array was designed to incorporate the physiologically relevant glycan structures that were identified in the glycan characterization of upper respiratory tissue. To characterize binding preferences, Influenza A viruses and recombinant HAs from various avian- and human-adapted IAVs were tested on the glycan array. Data-mining tools were then applied to glycan array data to identify patterns of structural features present in glycan receptors that differentiated avian *vs* human adapted HA binding specificity. This informatics-based approach showed that extension length (i.e. multiple lactosamine repeats) on the non-reducing end of sialic acid was important for human adapted HA binding whereas avian-adapted HAs demonstrated high affinity for α2→3 glycans and α2→6 glycans with short extensions [15]. These results were corroborated with molecular modeling analyses of co-crystal structures of HAs in complex with sialic acid receptors to reveal that specific topology (termed umbrella-like topology) adopted by extended α2→6 receptors differed from the topologies adopted by α2→3 glycans and short α2→6 glycans (termed *cone-like* topology) (Figure 1.11) [15]. Importantly, for glycans that adopt a cone-like topology, including both short α2→6 linked glycans and α2→3 linked glycans, the majority of the molecular contacts with the HA RBS are made by the trisaccharide Neu5Acα2→3/6Galβ1→3/4GlcNAc- motif. However, glycans that adopt an umbrella-like topology, such as α2→6 glycans composed of at least four monosaccharides, make additional contacts beyond the trissaccharide motif [58]. This detailed structural understanding of the glycan receptor and its topological presentation when bound to HA provides a unique insight into structural properties of HA-glycan interactions that govern specificity.

Based on these new findings, our lab wanted to develop a biochemical assay capable of assessing HA glycan specificity and affinity. Earlier studies investigating the


HA-LSTa Complex; Cone-like topology

HA-LSTc Complex; Umbrella-like topology

Figure 1.10 **Topology Influences HA-glycan specificity**. Interactions of HA with cone-like topology is characteristic of avian HA binding to α2-3 and short α2-6 glycans. In contrast, interactions of HA with umbrella-like topology is characteristic of human HA binding to long α2-6 glycans. Adapted from [58].

HA-glycan interaction had numerous limitations and often used assays with qualitative readouts, or employed glycans that only differed based on their terminal sialic linkage, and did not accurately capture the multivalent presentation of the glycan-protein interactions [37]. Based on our previous results, there was a need to develop an assay capable of quantitatively capturing HAs affinity for glycans that adopted cone- or umbrella-like topologies. Furthermore, given characteristic of glycan-proteins interactions, it was important to accurately capture the multivalent presentation of HA. To accomplish this, a dose-dependent glycan array was developed [41]. This assay employed glycans that could adopt cone-like topologies (denoted 3'SLN, 3'SLN-LN, 3'SLN-LN-LN, and 6'SLN) and those capable of adopting umbrella-like topologies (denoted 6'SLN-LN). Furthermore, this assay used an antibody-based pre-complexing strategy to enhance valence of HA presentation and ensure fixed avidity when comparing HAs [41]. Under these conditions the multivalent arrangement of pre-complexed glycans increased the binding signal, and improved the sensitivity of this assay (Figure 1.12). This binding assay provided a tool to quantitatively assess fine differences of binding specificity & affinity of HAs. Under this framework, it was confirmed biochemically that human-adapted HAs, such as HAs from pandemic strains, show high affinity for α2→6 glycans containing multiple lactosamine repeats (long α2→6, 6'SLN-LN), whereas avian-adapted HAs demonstrated high affinity for α2→3 glycans and α2→6 glycans with short extensions[15].

As discussed earlier, it was challenging to predict transmissibility properties of

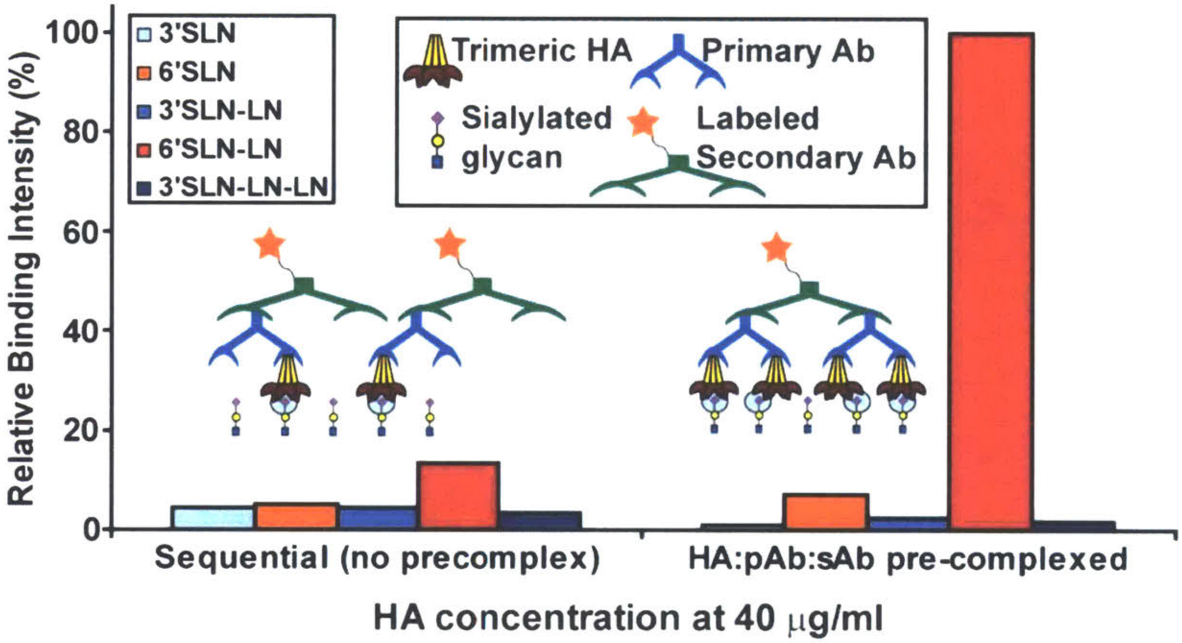


Figure 1.11 **Glycan Array Assay to Capture Multivalent HA-glycan Interactions**. Using either a sequential binding assay or precomplexation of HA units with primary and secondary antibodies, the glycan binding signal intensities were compared. Sequential assay favors the formation of HA:primary antibody:secondary antibody in 1:1:1 molar ratio as compared to precomplexing with primary and secondary antibodies before adding the glycan array favors a 4:2:1 ratio. The precomplexing strategy more accurately captures the multivalent presentation of HA. Adapted from [41].

HA reassortants that were mixed binders (i.e. agglutinated both α2→3 and α2→6 RBCs). For instance, HA reassortants containing NY18 showed inefficient respiratory droplet transmission, whereas Tx/91 shows efficient respiratory droplet transmission in ferret models [37]. Armed with these newfound insights into HA-glycan interactions, and a new biochemical assay to quantitatively assess HAs affinity for avian or human like glycan receptors that adopted the correct topology, the glycan binding properties of Tx/91 and NY18 were reexamined [41]. Indeed Tx/91 showed quantitatively stronger binding to glycans with an umbrella like topology (i.e. long α2→6 glycans 6'SLN-LN-), as compared to NY18. From these results, as well as the assessment of other HAs derived from pandemic or human-adapted IAVs, this analysis found that transmissibility was correlated with a quantitative shift in preference for binding glycans that adopt an umbrella-like topology [41].

Section 2: Integrated approaches to study cell surface glycans: Uncovering the role of heparan sulfate glycosaminoglycans in modulating cancer stem cell activity through regulation of cell-microenvironment interactions

Motivation and Objectives

Cancer stem cells are defined as a subpopulation of cancer cells functionally defined by their ability to initiate new tumors [64]. Furthermore, and in accordance with their name, they can also self-renew and undergo differentiation to give rise to the various diverse cell types that are not tumor-initiating that comprise the tumor [65,66]. One key implication of this cancer stem cell model is that CSCs, similar to adult stem cells, are hierarchically organized where cancer stem cells reside at the apex of the hierarchy. Furthermore, cancer stem cells are characterized by their resistance to chemotherapy and radiation, and are qualified to seed metastasis or regrow tumors after relapse [67]. Given the role that CSCs play in tumor progression, it is possible that therapeutically targeting CSCs in tumors could lead to a durable clinical response or prevent metastatic disease. In this context, there is great interest in uncovering the CSC-specific molecular pathways that regulate CSC maintenance and activity.

Interactions between CSCs and their microenvironment are important for CSC behaviors such as invasion, migration, resistance to anoikis, chemotherapeutic resistance and tumor initiation [68]. The CSC niche is highly complex and is composed of numerous molecular players including: Hormonal signals, paracrine and autocrine signals such as morphogens/cytokine/growth factors, extracellular matrix molecules that directly interact with CSC [69–71]. Additionally, cellular player such as activated stroma, immune cells, or surrounding tumor cells create and modify the microenvironment and direct CSC activity [64]. Preclinical evidence suggests that targeting CSC-microenvironment interactions in cancer might reduce the CSC number or activity, thereby improving disease outcomes [72]. It remains unknown whether or not targeting CSC-microenvironment will yield clinical benefit and numerous clinical trials are underway testing this hypothesis [73]. Thus, the microenvironment of CSCs and their interactions represent fertile ground for the discovery of new therapeutic targets.

HSGAGs are abundant on the cell surface and at the cell-ECM interface where they mediate interactions between cells and their microenvironment. Despite knowledge that HSGAGs play a critical role in cell-microenvironment interactions, the function of HSGAGs in cancer stem cell activity is largely unknown. The objective of section 2 is to elucidate the role that HSGAGs, as regulated by SULF1, play in modulating breast cancer stem cell activity (Chapter 8).

Glycosaminoglycans and the cellular microenvironment

In recent years, strong evidence has emerged demonstrating that the niche, or microenvironment, in which a cell resides plays a dominant role in cell phenotype and function [69–71]. The term microenvironment is used to described the totality of molecular signals and cellular players which contribute to a cell's biological function. The microenvironment includes extracellular matrix proteins, soluble factors such as hormones, growth factors, enzymes, and cellular players such as stromal cells or juxtacrine signaling cellular neighbor [70,74,75].

Despite the growing appreciation for the important biological function of the microenvironment, certain components in the microenvironment, namely glycosaminoglycans have largely gone unstudied. Glycosaminoglycans (GAGs) can exists as free polysaccharides or can be attached to a protein backbone[76]. GAGs are primary components of the extracellular matrix and the cell surface where they play various roles including, cell adhesion, sequestration of soluble protein, providing turgor of soft tissues, and many more [4,77,78].

Broadly, GAGs are linear polysaccharides composed of a repeating disaccharide building block of uronic acid (either β-D-glucuronic acid or α-L-iduronic acid) which is linked (either 1→3 or 1→4) to an amino sugar (N-acetyl glucosamine or N-acetyl galactosamine) [4,79]. There are four main classes of GAGs which differ based on their disaccharide repeat and chain length. The major categories of GAGs include Heparin/Heparan Sulfate (HS or HSGAGs), Chondroitin/Dermatan Sulfate (CS/DSGAGs), Hyaluronic acid (HA), and Keratan Sulfate (KS). The disaccharide building block for each class is shown in Figure 1.13. The focus of the work in this thesis is HSGAGs, and thus CS/DSGAGs, HA or KS will not be discussed further.

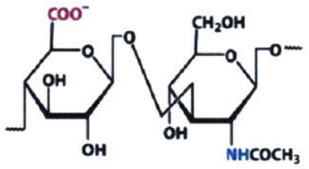
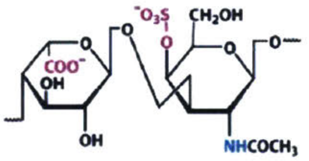
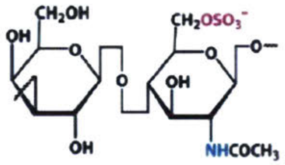
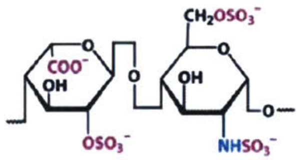
Main Classes of GAGs	Disaccharide Repeat Unit	Number of Disaccharides	Examples of Tissue Distribution in Mammals
Hyaluronan 	Glucosamine linked to glucuronic acid	250-25,000	Skin, skeletal tissues, synovial fluid and the vitreous humor of the eye
Chonroitin/Dermatan Sulfate 	Galactosamine linked to uronic acid	Average of 40	Cartilage, brain, connective tissue, fibroblasts, neural cells, endothelial cells, lymphocytes and myeloid cells
Keratan Sulfate 	Glucosamine linked to galactose	≤ 50	Bone and cartilage
Heparin/Heparan Sulfate 	Glucosamine linked to uronic acid	20 to 200	Mast cells (heparin), every type of cells (heparan sulfate)

Figure 1.12 **GAG classification and general description.** Table shows the structure of the disaccharide repeat unit, number of disaccharides that make up a chain, and examples of tissue distribution of the 4 main classes of GAGs. Adapted from [79]

HSGAG structure

HSGAGs are the most structurally diverse member of the GAG family. HSGAGs are typically found attached to a protein, termed proteoglycans[80]. Despite being considered a post transitional modification, HSGAGs are known to exert function independent of the protein backbone and liberated HSGAGs, such as heparin, have potent bioactivity. Structurally, HSGAGs are linear polymers, composed of a repeating disaccharide unit of uronic acid that is 1→4 linked to N-acetyl glucosamine, strung together in chains ranging from 20-200 disaccharide units [81]. HSGAG chains can be chemically modified (sulfated or acetylated) at various positions on the disaccharide, or modified through epimerization. For instance, epimerization at the C-5 sugar in the uronic acid component results in either β-D-glucuronic acid or α-L-iduronic acid. Furthermore, sulfation can occur at the C-2 position of the uronic acid (denoted 2-O-sulfation (2S)), the C-3 and C-6 positions of the glucosamine (3-O and 6-O-sulfation (3S, 6S)). Additionally, N-sulfation (N-S) or N-acetylation (N-Ac) can occur at the C-2 of the glucosamine (Figure 1.14) [5]. In total, there are 24 uniquely sulfated or acetylated disaccharide repeat structures, per uronic acid isomer. Thus, there are 48 possible unique disaccharides in total. From a combinatorial perspective, compared to other linear biopolymers such as DNA (4 bases) or protein (20 amino acids), HSGAG (48 disaccharides) are the most structurally diverse and information dense linear biopolymers in nature. In addition to the rich structural diversity, the biosynthesis of HSGAGs in non-template driven [82,83]. This results in heterogeneity at the level primary 'sequence' of HSGAGs and chain length. Taken together these features of HSGAGs have posed enormous challenges in deciphering structure-function relationship of HSGAGs.

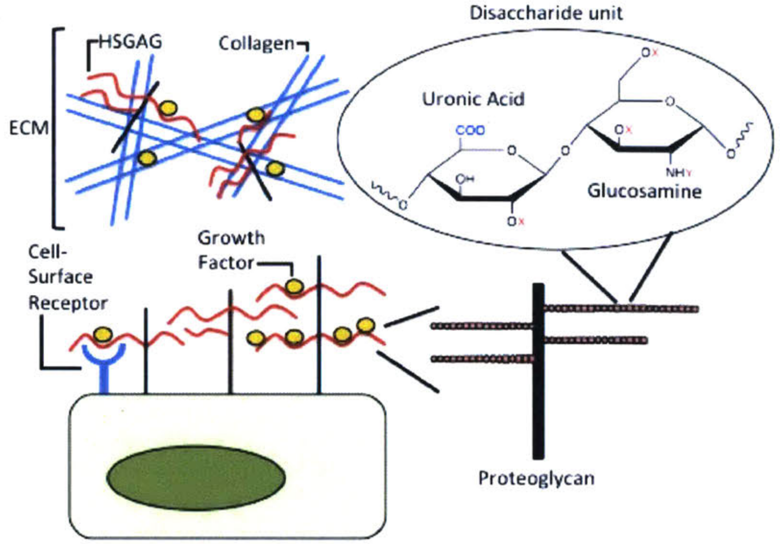


Figure 1.13 **Structure and biology of heparan-sulfate glycosaminoglycans.** HSGAGs exist at the cell surface and in the ECM, where they are attached to a proteoglycan core. HSGAGs are complex, linear polysaccharides comprised of repeating disaccharide unit uronic acid linked to a glucosamine. Each disaccharide unit can be sulfated at the 2-O position of uronic acid and the 3-O and 6-O position on glucosamine (X=sulfation). The N-position of glucosamine can acetylated, sulfated or unmodified (Y=acetylation or sulfation). The diversity of chemical sequences in polysaccharide chains enables HSGAGs to bind and modulate the activity of growth factors, chemokines, and enzymes both at the cell surface and in the ECM. Adapted from [5].

Biosynthesis of HSGAGs

The heterogeneity and chemical complexity of HSGAGs originates through their complex biosynthesis. HSGAGs are synthesized attached to a proteoglycan core. Thus, on the cell surface HSGAGs chains are presented in the context of proteoglycans, examples of these proteoglycans include syndecan, glycpican and perlecan [80].

As mentioned above, the biosynthesis of HSGAGs is non-template driven and instead involves the co-ordination of multiple enzyme and enzyme complexes in the golgi. There are three key major steps of HSGAG biosynthesis: i) the synthesis of the tetrasaccharide linkage sequence and attachment to the proteoglycan core, ii) extension of the HSGAG chain through co-polymerization of glucuronic acid and N-acetylglucoseamine, and iii) chemical modification HSGAG chains [described below Figure 1.15 [4,79]].

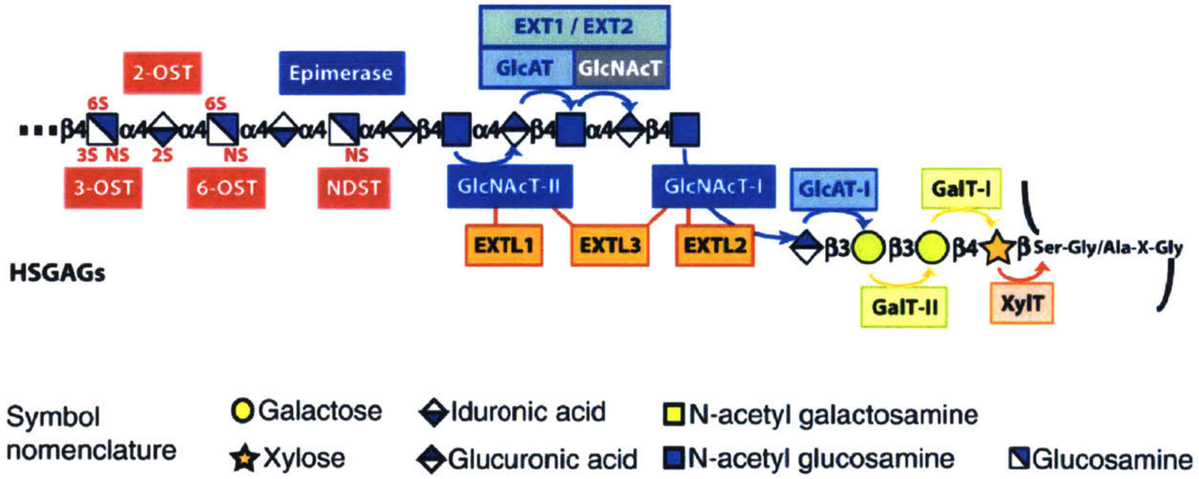


Figure 1.14 **Biosynthesis of GAGs**. The biosynthesis of HSGAGs, starting from the core protein, including glycosyltransferases and sulfotransferases involved, is illustrated using the symbol nomenclature for glycans. Adapted from [4]

HSGAGs biosynthesis begins with synthesis and attachment of the tetrasaccharide linker, glucuronic acid-galactose-galactose-xylose-to the proteoglycan core at the consensus sequence (Ser-Gly/Ala-X-Gly (where X is any amino acid)), yielding GlcAβ1→3Galβ1→3Galβ1→4Xylβ1→O-(Ser)-. This process is regulated by the coordinated action of four enzymes: xylosyl transferase, β4-galactosyl transferase (GalT- I), β3-galactosyl transferase (GalT-II), and β3-GlcA transferase (GlcAT-I). Next, the HSGAG chains are elongated through alternating addition of GlcA and GlcNAc by a multidomain glycosyltransferase, EXT1 and EXT2 [4,79].

The final step of HSGAG biosynthesis is chemical modification. These enzymes imbue the nascent heparan sulfate chains with O-sulfation at the 2-O, 3-O, and 6-O position, N-sulfation or acetylation, and epimerization of glucuronic acid to iduronic acid. This process involves the coordination of multiple enzymes that can act in concert, specifically it involves the sequential activity of N-deacetylase, N-sulfotransferase (NDST), C-5 Epimerase, 2-O sulfotransferase (2-OST), 3-O sulfotransferase (3-OST) and 6-O sulfotransferase (6-OST). Briefly, first NDST cleaves N-acetyl groups from GlcNAc and adds N-sulfation. Then, Epimerase converts glucuronic acid to iduronic acid. Next, 2-OST sulfates glucuronic acid/iduronic acid at the C-2 position, with preference for iduronic acid. The final modifications are performed by 3-OST and 6-OST, which sulfate the C-3 and C-6 of glucosamine. For each of the above enzymes several isoforms have been identified, including four NDST isoforms, six different 3-OSTs, and three 6-OSTs[84]. Of note, these isoforms are often expressed in a tissue specific manner, suggesting tissue specific regulation of HSGAGs [4,84].

Similar to the way proteins are described by their linear primary amino acids, HSGAG primary sequence can be described by their sulfation/modification pattern. The biosynthetic process of HSGAGs generates enormous primary sequence diversity. Furthermore, a clustered organization has been observed, where HSGAGs can be characterized by N-Sulfated domains, NS/NAc transition domains and NAc domains based on the sulfation of glucosamine [85]. Indeed, HSGAGS derived from various biological contexts show different sequences, and domain organization characteristics. For instance, HSGAGs present in cell surfaces are termed heparan sulfate, which is distinct from heparin. Heparin is synthesized attached to serglycan and is stored in intracellular granules of mast cells, acting as a reservoir for proteases. Structurally, heparin is highly sulfated and composed of the trisulfated dissacharide $-[I_{2S}-H_{NS,6S}]_{(n)}-$. In contrast, heparan sulfate can be attached to a diverse array of proteoglycans and is less sulfated than heparin[4].

Biological function of HSGAGs

HSGAGs regulate numerous key biological functions at the cell-microenvironment interface and in the ECM. HSGAG play three major functions; i) they act as co-receptors for growth factors/chemokines to facilitate oligomerization or receptor-ligand complex formation, ii) create gradients through affinity-based localization of HS binding components, and iii) store or sequester growth factors & enzymes in at the cell surface and in the ECM (Figure 1.14) [86]. Examples of each are described below:

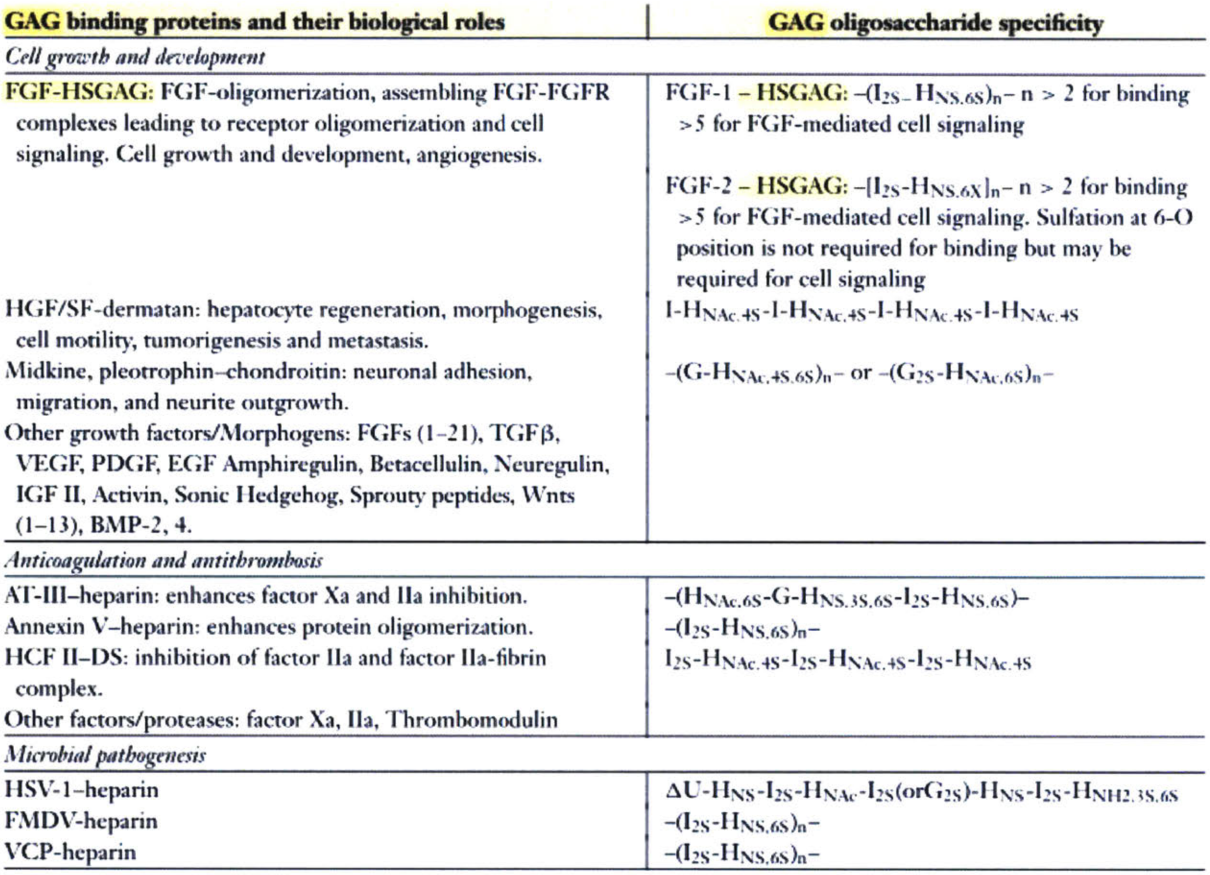
i) HSGAGs interact with proteins at the cell surface and in the ECM where they facilitate many receptor-ligand interactions. In the case of FGF-2, heparin is a key regulator of FGF signaling activity. Structural and biochemical studies have revealed that HSGAGs interact directly with FGF-2 to facilitate dimerization. Additionally, HSGAGs facilitate the formation of FGFs active signaling complex, or ternary complex. This ternary complex is composed of FGF dimers, bound to FGFR dimers, also bound to HSGAGs in a 2:2:2 ratio [87].

ii) HSGAG also acts as a key player in the formation of morphogen or growth factor gradients. Specific HSGAGs structures in the ECM can interact with morphogens or growth factor to modulate their diffusion. For example, an extracellular sulfatase (which removes 6-O sulfated sugars post-synthetically) is a critical regulator of WNT, which influences embryo patterning [88].

iii) HSGAGs are also known to act as a reservoir for factors in the cell surface or ECM. For instance, in a cell surface context, WNT preferentially binds to 6-O sulfated HSGAGs on the cell surface rather than to its receptor, frizzled. Upon expression of extracellular sulfatases, 6-O-sulfates are cleaved and WNT ligands are mobilized enabling their receptor-ligand interaction [89]. Similarly, in an ECM context, HSGAG in the ECM sequester and store bFGF, which can be rapidly mobilized through degradation of the ECM or HSGAGs in the ECM [78].

HSGAG protein interactions

HSGAGs exert their function through their interaction with proteins. HSGAG-protein interactions result in various functional effects including: ligand immobilization to generate storage depot or gradients, protection of proteins from degradation, oligomerization of growth factor, and induction of conformational changes to enhance interactions or induce signaling [4,5,79]. Initially, the interaction between HSGAGs and proteins was assumed to be driven solely by electrostatic association. However, it is now appreciated that HSGAG-binding proteins show a high degree of sequence specificity and selectivity [90]. Indeed, many studies have attempted to understand the heparin binding specificity of HSGAGs. Various level of structural detail can be characterized including: correlational relationships between protein binding & function and HSGAGs disaccharide composition (i.e. WNT is negatively regulated by 6-O-endosulfatases, where WNT binds 6-O-sulfated HSGAGS with high affinity[89]), elucidation of the linear glycan sequence of the HS binding motif (i.e. AT-III binds (HNAc,6S-G-HNS,3S,6S-I2S-HNS,6S) [90]), and detailed structure information such as topological and conformational requirements for binding (i.e. FGF-2 binds a repeat of $[I_{2S}H_{NS,62}]_n$ which contains a kink motif [H-I-H] enabling high affinity HSGAG-FGF-2 binding through optimal van der Waals interactions [91,92]). A representative list of GAG binding proteins and their biological function along with their oligosaccharide specificity is included in Figure 1.16.



GAG binding proteins and their biological roles	GAG oligosaccharide specificity
Cell growth and development	
FGF-HSGAG: FGF-oligomerization, assembling FGF-FGFR complexes leading to receptor oligomerization and cell signaling. Cell growth and development, angiogenesis.	FGF-1 – HSGAG: $-(I_{2S-}H_{NS.6S})_n-$ n > 2 for binding >5 for FGF-mediated cell signaling FGF-2 – HSGAG: $-[I_{2S}\text{-}H_{NS.6X}]_n-$ n > 2 for binding >5 for FGF-mediated cell signaling. Sulfation at 6-O position is not required for binding but may be required for cell signaling
HGF/SF-dermatan: hepatocyte regeneration, morphogenesis, cell motility, tumorigenesis and metastasis.	$I\text{-}H_{NAc.4S}\text{-}I\text{-}H_{NAc.4S}\text{-}I\text{-}H_{NAc.4S}\text{-}I\text{-}H_{NAc.4S}$
Midkine, pleotrophin–chondroitin: neuronal adhesion, migration, and neurite outgrowth.	$-(G\text{-}H_{NAc.4S.6S})_n-$ or $-(G_{2S}\text{-}H_{NAc.6S})_n-$
Other growth factors/Morphogens: FGFs (1–21), TGFβ, VEGF, PDGF, EGF Amphiregulin, Betacellulin, Neuregulin, IGF II, Activin, Sonic Hedgehog, Sprouty peptides, Wnts (1–13), BMP-2, 4.	
Anticoagulation and antithrombosis	
AT-III–heparin: enhances factor Xa and IIa inhibition.	$-(H_{NAc.6S}\text{-}G\text{-}H_{NS.3S.6S}\text{-}I_{2S}\text{-}H_{NS.6S})-$
Annexin V–heparin: enhances protein oligomerization.	$-(I_{2S}\text{-}H_{NS.6S})_n-$
HCF II–DS: inhibition of factor IIa and factor IIa-fibrin complex.	$I_{2S}\text{-}H_{NAc.4S}\text{-}I_{2S}\text{-}H_{NAc.4S}\text{-}I_{2S}\text{-}H_{NAc.4S}$
Other factors/proteases: factor Xa, IIa, Thrombomodulin	
Microbial pathogenesis	
HSV-1–heparin	$\Delta U\text{-}H_{NS}\text{-}I_{2S}\text{-}H_{NAc}\text{-}I_{2S}(\text{or}G_{2S})\text{-}H_{NS}\text{-}I_{2S}\text{-}H_{NH2.3S.6S}$
FMDV-heparin	$-(I_{2S}\text{-}H_{NS.6S})_n-$
VCP-heparin	$-(I_{2S}\text{-}H_{NS.6S})_n-$

Figure 1.15 **A table of GAG binding proteins**: their biological function and oligosaccharide specificity adapted from [4]

HSGAGs and their role in Cancer

Given the numerous important biological roles that HSGAGs play at the cell-microenvironment interface, it is not surprising that they play a critical role in tumor progression. HSGAGs regulate multiple stages of tumor progression including tumor growth, angiogenesis, and metastasis [5]. In fact, specific sequences of HSGAGs can differentially regulate tumor growth and metastasis. In a study performed in our laboratory, tumors were treated with bacterial heparinases (Hep I vs Hep III) which have distinct substrate specificities [93]. This study revealed that Hep I treatment (which cleaves highly sulfated regions of HSGAGs) promoted tumor growth whereas, HepIII treatment (which cleaves unsulfated regions of HSGAGs) inhibited tumor growth and suppressed metastasis [93]. This study provides an elegant demonstration of the diverse information encoded in HSGAGs structure. This further highlight the need to understand structure-function relationship, as biological function is regulated by HSGAGs in a sequence specific manner.

HSGAG sequence alone does not dictate function, and the microenvironment context greatly influences the tumor phenotype. For instance, two extracellular endosulfatases, SULF1 and SULF2, have emerged as key regulators of growth factor signaling, demonstrating critical roles in development, tumor growth and metastasis [5,88,94]. SULFs remove the 6-O sulfate (6-O-S) modification of the glucosamine with a preference for the tri-sulfated disaccharides and have been cited as having both tumor suppressing and oncogenic functions. For example, in multiple myeloma, the experimentally induced expression of SULFs reduced tumor growth *in vivo*, implicating SULFs as tumor suppressors. However, In the context of lung cancer, SULFs possessed oncogenic activity, as SULF knockdown decreased cell growth, tumor formation, and cell migration.

These seemingly paradoxical functions can be explained by the nature of the growth factor-HSGAG interaction [95,96]. 6-O-Sulfation of HSGAGs is known to impinge on a variety of growth factor signaling pathways. For instance, In the case of FGF-2 signaling, 6-O-S modified HS is required for the formation of the active FGF-2:HSGAG:FGFR1 signaling complex [97]. Acting in the opposite direction are WNT ligands, which demonstrate high-affinity binding to 6-O sulfated HS; this binding prevents functional interaction of WNTs with their cognate receptor, Frizzled. In this context, high expression of SULFs would reduce the overall 6-O-S modification of HSGAGs, thereby liberating previously sequestered WNTs and enabling binding to Frizzled [89]. Indeed, in multiple myeloma experimental expression of SULFs reduced the formation of FGF:FGFR1 ternary complexes. In the context of the multiple myeloma microenvironment, FGF acts a positive regulator of tumor growth, giving rise to the observed tumor suppression phenotype. However, In the context of lung cancer microenvironment, the decrease in WNT signaling driven by SULF knockdown, led to the decreased cell growth, tumor formation, and cell migration [98]. These finding suggest that phenotype resulting from tumor cell-microenvironment interactions is both HSGAG sequence and microenvironmental context dependent.

Lastly, HSGAG are a viable target for cancer therapy. Often HSGAG dependant signaling can be targeted through the inhibition of several extracellular HSGAG modifying enzymes, such as heparanase (HPSE) [99]. In many human tumors the enzyme HPSE is upregulated [100,101]. HPSE, is an extracellular β-D-glucuronidase which is involved in angiogenesis [99]. It cleaves heparan sulfate chains to mobilize pro-angiogenic factors stored in the ECM. Of note, high expression of HPSE has been linked to poor prognosis [100]. Due to HPSE's role in cancer progression, there has been much interest in using heparin or heparan sulfate mimetics to target HPSE activity in tumors [102,103]. The utility of sulfated oligosaccharides to treat tumors has been demonstrated in a phase II clinical study for hepatocellular carcinoma [104]. Numerous preclinical and clinical trials are underway to develop anti-metastatic, anti-angiogenic therapies targeting HSGAGs modifying enzymes [5].

Structural characterization of HSGAGs

Towards decoding structure-function relationships, a first key step is to characterize the sequence or structure of HSGAGs. Structural analysis of HSGAGs is challenging due to a high degree of structural complexity, heterogeneity, and low abundance. To address these challenges, a unique toolkit comprised of HSGAG degrading enzymes and high sensitivity analytical tools has been previously developed.

HSGAG degrading enzymes from bacteria, such as Hep I, II, & III have proved invaluable tools in dissecting HSGAG structure function relationships [105]. Each of these heparinases cleaves different HSGAG sequences through a lytic enzyme mechanism. In general, Hep I acts on highly sulfated domains of HSGAGs while Hep III cleaves lowly sulfated domains. However, Hep II activity is not dependent on HSGAG sulfation [105]. Conveniently, the lyase activities of Hep I, II, III imbue the uronic acid with a Δ4,5 unsaturated bond which is a chromophore enabling label-free detection using UV (at 232nm).

From an analytical perspective, HSGAGs possess many features, such as their strong negative charge, large size (relative to branched glycans), and the abundance of isobaric structural isomers makes sequencing intact HSGAG chains intractable [4]. However, using one or more HSGAG degrading enzymes, a controlled digestion of HSGAGs can yield fragments that can be more easily characterized using analytical techniques, such as mass spectrometry [106,107]. Furthermore, exhaustive depolymerization of HSGAGs using a collection of hep I, II, and III can reduce an HSGAG chain into disaccharide units [108]. Using capillary electrophoresis, the sulfated isomers can be resolved, yielding the relative abundance of mono-, di- and tri-sulfated disaccharide as well as resolution of positional isomers [106,107,109]. There is no one technique, capable of measuring all the structural attribute of HSGAGs. As such the sequencing of HSGAGs requires integration of multiple orthogonal analytical measurements, including: NMR which yields insight into monosaccharide identity and linkage, Capillary electrophoresis which measures the disaccharide composition, and MALDI and ESI mass spectrometry which provide insight into chain length and mass-composition relationships [4].

Measurements of HSGAG structure alone are necessary but not sufficient to decode structure-function relationship. Building sequence or composition to function relationships requires integrated insight from functional assaying in cell or organism model systems, as well an understanding of the structural basis for HSGAG-protein interaction affinity and specificity [93,110,111]. Furthermore, assessing HSGAGs biological activity requires the ability to perturb heparan sulfate structure or its interactions with heparin binding protein. In the context of HSGAGs, this is done using HSGAG degrading enzymes, chemical inhibitors, or manipulation of HSGAG biosynthesis genes [93,104,111]. Ultimately, using this integrated approach (colloquially termed a functional glycomics approach) it is possible to decode structure-function relationships of HSGAGs [4].

Cancer stem cells

Cancer stem cells are a subpopulation of cancer cells functionally defined by their ability to initiate new tumors [64]. Furthermore, and in accordance with their name, they can also self-renew and undergo differentiation to give rise to the various diverse cell types that comprise the tumor [65,66]. One key implication of the cancer stem cell model is that CSCs in tumors, similar to adult stem cells in tissue, are hierarchically organized where cancer stem cells reside at the apex of the hierarchy. Furthermore, cancer stem cells are characterized by their resistance to chemotherapy and radiation, and they are known to be key contributors to metastasis and relapse [67]. Given the role that CSCs play in tumor progression, it is possible that therapeutically targeting CSCs in tumors could lead to a durable clinical response or prevent metastatic disease. In this context, there is great need to uncover the CSC-specific molecular pathways that regulate CSC maintenance and activity.

Cancer stem cells were first identified based on the experimental demonstration that a minority subpopulation of primary human acute myeloid leukemia initiated tumors in immunocompromised mice at a much higher frequency compared to the bulk cell population [112]. Furthermore, this tumor initiating subfraction exhibited cell-surface marker phenotypes similar to that of a hematopoietic stem cell. CSC population have been isolated and identified in many tumor subtypes including, breast [66], colon [113], brain [114] and many more [115]. Still today, CSCs are functionally defined by similar measures. In fact, the most robust measure used to identify cancer stem cells in the limiting dilution assay (LDA). In LDA tumor initiation frequency is measured following implantation of log fold dilutions of tumor cells into to a NOD/SCID mouse. Additionally, CSCs possess several other characteristic traits including resistance to anoikis, and increased resistance to chemotherapy and radiotherapy [64,65,116]. Despite our ability to identify CSCs, the molecular mechanisms that underlie their functional properties remain largely unknown.

Identification and isolation of CSCs is central our ability to study CSCs biology. Cancer stem cells possess unique molecular features that can be used for their identification and isolation by FACS. Often CSCs are characterized by the semi-quantitative expression of one of more cell surface markers. For instance, cancer stem cells from breast tumors are characterized by high expression of CD44, with concomitantly low or absent CD24 expression, thus the CSC population is denoted CD44^{hi}, CD24$^{low/-}$ [66]. To date, various cell surface markers have been identified for CSCs arising from a number of tumor tissue types (i.e CD34, CD44, CD90 are reported CSCs markers for hematologic malignancies, breast, and brain, respectively [66,112,114]). Of note, there tends to be a high degree of similarity between surface markers that identify CSCs and those identifying normal tissue stem cells. Additionally, there are some non-cell surface marker methods to identify CSCs. For instance, populations of cells that can efflux hoeschst 33342 dye, termed 'side populations' are enriched for CSCs. Importantly, there is no set of markers that can uniformly identify CSCs [117].

During malignant progression, epithelial cancer cells are thought to acquire mesenchymal traits [118]. The cell- biological program that facilitates acquisition of these traits is known as the epithelial to mesenchymal transition (EMT). During passage through an EMT, epithelial cells lose their differentiated characteristics such as cell–cell adhesion and lack of motility, and acquire the traits of mesenchymal cells which include migration or invasion and elevated resistance to apoptosis. The activation of an EMT program in cancer cells imbues them with stem-like properties, and is thought to enable tumor cells to progress through the metastatic cascade [69,116,119]. Importantly, CSC are known to possess several mesenchymal traits and these traits are thought to give rise to their characteristic TIC behaviors such as invasion, migration, resistance to anoikis, chemotherapeutic resistance and tumor initiation. CSCs are thought to reside in a partial EMT'd state, where they possess features of both epithelial and mesenchymal cells [120].

As EMT is a key program for generating CSCs, an important focus of the field is on elucidating the EMT-associated signals that regulate CSC maintenance and function. EMT is regulated by various signals from the cell microenvironment, such as, ECM components (such as collagen or laminin) and secreted factors such as TGFβ and WNTs[69]. These signals induce the expression of the EMT-transcription factors, SNAIL, TWIST and ZEB, to drive the suppression of the epithelial genes and the expression of the mesenchymal genes [118,120,121]. There are many additional molecular players involved in EMT and the extensive collaboration between them during signal transduction, these have been reviewed elsewhere [121] and will not be discussed further as they are outside the scope of this thesis.

The association between CSCs and their regulation of EMT provides fertile ground for the discovery of potential therapeutic interventions. In the context of targing EMT two forms of therapeutic intervention are compelling: i) targeting EMT specific signaling networks which control tumor initiation, invasion, resistance to anoikis, or ii) pushing cells out of their EMT'd state (or inducing an MET). The latter has been difficult to achieve, but holds strong therapeutic promise.

Thesis objectives

The two major goals of this thesis are to i) further develop and improve integrated approaches to study glycans and glycan-binding proteins, and ii) leverage these approaches to uncover new biological roles of glycans and GBPs in disease. Thus, this thesis is comprised of 5 major objectives:

Part 1: Integrated approaches to study glycan binding proteins.

Objective 1: Implement an integrated approach that incorporates an inter-residue interaction network tool to study glycan-GBP interactions. Investigate the ability of this tool to efficiently identify key functional residues in the glycan binding site of a model GBP (Chapter 2).

Objective 2: Apply the integrated experimental framework developed in Objective 1, to the study of influenza A hemagglutinin and HA-glycan interactions, with the goal of uncovering insights to improve pandemic risk assessments of IAVs from non-human hosts (Chapter 3-6). Specific sub-objectives:

e. Identify the structural determinants for naturally evolving H5N1 HA to switch its receptor specificity from avian to human (Chapter 3)

f. Determine the physiologic glycan receptor binding properties of a 2013 outbreak H7N9 HA and identify the mutations required to enable binding to human glycan receptors. (Chapter 4)

g. Develop bioinformatics tools and experimental methods capable of measuring antigenic properties of HA. Leverage these tools to identify avian and swine-adapted H3s that could re-emerge into the human population and potentially cause a pandemic (Chapter 5).

h. Characterize the glycan binding specificity of an HA isolated from the 2011 New England harbor seal H3N8 (Chapter 6)

Part 2: Integrated approaches to study glycans and their biological function

Objective 3: Implement an integrated approach to characterize cell surface glycans. Investigate the ability of this approach to characterize the N- and O-linked glycomics changes associated with metastatic progression in a cell line derived from a mouse model of multistage lung adenocarcinoma (Chapter 7).

Objective 4: Elucidate the role that HSGAGs, as regulated by SULF1 play in modulating breast cancer stem cell activity (Chapter 8).

Objective 5: Propose an analytical framework for the development and quality control of immunomodulatory therapeutics derived from glycan-containing dust extracts isolated from dairy cattle barns (Chapter 9)

Chapter 2 : Implementation of a residue interaction network analysis tool to study glycan-protein interactions

Summary and Significance

Glycan-binding proteins (GBPs) interpret the chemical information encoded by glycans. Thus, in order to understand the function of glycans, it is important to study glycan-GBP interactions, specifically, the structural determinants that govern specificity and affinity. In order to study these structural determinants, an experimental approach that integrates biochemical assays, functional assays, and structural analysis is required. Importantly, there is still a great need to develop new tools that improve this integrated experimental approach. In this chapter, I describe the development and implementation of a unique experimental approach that improves the identification of functional residues in the glycan binding site on GBPs. I focus on applying a new tool to this integrated analysis, namely an inter-residue network analysis tool, called SIN (Significant Interaction Network) recently developed in Ram Sasisekharan's lab. SIN adds a complementary view of protein structure where proteins are represented as network graphs. Amino acids are represented as nodes and their interactions are connecting edges. The network can be analyzed qualitatively and quantitatively to identify low-, moderate- and highly-networked amino acids in proteins. Previous work performed using amino acid network analysis suggested that it may be possible to use network analysis for ligand binding site prediction. Thus, I explore if our network analysis tool could be used to identify important functional residues in the glycan binding site of a model GBP, bFGF. In this chapter of my thesis, I demonstrate that: i) SIN can identify known functional residues on the bFGF protein, ii) that SIN-guided mutagenesis can elaborate critical residues in the heparin binding site of bFGF, and iii) preliminary results show that SIN has helped identify a new highly networked region of unknown function on bFGF. These results suggest that SIN could be applied to the study of other glycan-binding proteins to help elucidate additional structural determinants that are critical for glycan-GBP interactions.

Introduction

Glycans play numerous important biological roles such as modulating receptor signaling, cell growth, cell-ECM interactions, and host-pathogen interactions [122]. The method by which glycans exert their function is often through their interaction with glycan-binding proteins. Thus, studying glycan-GBP interactions is critical to uncover the biological function of glycans. However, studying GBPs is very challenging; GBP-glycan interactions are often low affinity (uM-mM), and GBPs often bind many diverse but structurally related glycans with varying affinity. Furthermore, specific, high affinity interactions between glycans and GBPs are achieved through their multivalent presentation and are driven by avidity. Finally, structural features of the glycan such as conformation and topology play a crucial role in glycan-GBP interaction. Due to these complexities, determining structure-function relationships can be very challenging. Consequently, when studying glycan-protein interactions it is often necessary to obtain a detailed atomic level view of their interaction. Additionally, structural analyses must be integrated with experimental measurements to enumerate structure-function relationships of glycans and their interacting GBPs.

Numerous strategies have been applied to identify the structural determinants of glycan-protein interactions. These strategies often employ biochemical assays to assess specificity & affinity (i.e. glycan array binding assays), methods to generate a molecular level view of interactions (i.e. NMR or x-ray crystallography of glycan-protein complexes), computational tools to explore glycan-protein interfaces (i.e. visualization tools, docking software, molecular dynamics simulations), and functional readouts [87,91,123,124]. Additionally, analytical tools such as selective labeling and molecular biology approaches like random or structure-guided mutagenesis are needed for identification and confirmation of the structural determinants of a glycan-GBP interaction [125]. Currently, these integrated approaches are expensive, time consuming and require many iterations between structural models and experiments. Thus, there is still a great need to develop new computational tools to improve this process and improve prediction of key structural determinants on proteins. Ideally, a computational tool would provide information about the glycan-protein interaction, while reducing the amount of time-consuming, resource-intensive experimentation required.

One approach that has recently gained popularity is transforming 3D protein structures into amino acid networks [126–128]. Proteins are linear polypeptides of amino acids folded into a three-dimensional protein structure. Importantly, in the context of a folded structure these amino acids do not exist in isolation; they make a number of inter-atomic interactions with protein backbone, amino acid sidechains and solvents. These can be captured using network analysis where each residue is depicted as a node, and each interaction as an edge. A network view of a protein structure, can allow for the study of topological information as well as global connectivity of residues in the protein molecule. Applications of network approaches to study proteins can be used to investigate protein folding, allosteric communication analysis, interactions within protein complexes, and hot-spot residues [126,128]. It has also been suggested that network analyses might help identify functionally important residues in a protein-ligand interaction [127]. Thus, it is possible that this tool could have strong utility in the study of structural determinants of glycan-GBP interactions.

Recently, a new network analysis tool to study inter-residue interaction (hereafter SIN) was developed in Ram Sasisekharan's lab (see "methods" section for details) [129]. SIN enables the investigator to explore the local network of an individual amino acid. For each residue, a SIN score is calculated which takes into account all the atomic interactions, such as hydrogen bonds, disulfide bonds, pi-bonds, polar interactions, salt bridges and van der Waals. SIN yields two key pieces of information: i) a qualitative description of a residue's network, where a residue is depicted as a node and an interaction type as an edge, and ii) a quantitative "score" that is determined from the integration of all the interaction types which can be used to compare residues within a protein (Figure 2.2a). Previously, our lab's SIN tool was used to explore the link between antigenic regions on hemagglutinin proteins (which are subjected to mutational pressure) and the receptor binding site. In that study, using SIN, the authors identified that amino acids in the influenza viral hemagglutinin with a high network score are structurally constrained to mutate [129]. Furthermore, this work led to new hypotheses about how mutations that occur in antigenic sites of influenza might give rise to changes in its receptor binding properties. This was an important interaction to investigate, as previous experimental studies have demonstrated that antigenic escape mutants, which arise in pre-vaccinated hosts, drive an increase in receptor binding affinity [130].

SIN analysis was shown to: i) identify residues that are constrained to mutate, ii) elucidate the structural effects of mutations, and iii) show how various intraprotein domains interact. Other network analysis tools have shown that residues with a high degree of closeness centrality and surface accessibility may represent structurally conserved residues which could have important protein function [127,129]. In this study, I aim to investigate the ability of SIN to identify functional residues present in the glycan binding site of GBPs. In the context of this thesis work, this tool may be useful in elucidating the structural determinants of glycan-protein interactions. These structural determinants are critical to uncover in order to understand glycan-protein selectivity, specificity, affinity. Here, using a well-studied model system protein, bFGF, I explore the utility of our SIN tool to elucidate the structural determinants of heparin-bFGF interactions.

Experimental Methods

Reagents

Chemical reagents for buffers were purchased from Sigma Aldrich. A mouse monoclonal anti-6xHis antibody was purchased from Abcam (AB18184); Goat Anti Mouse-HRP secondary antibody was purchased from Santa Cruz (SC-2005). Amersham ECL prime western blot detection reagent was purchased from GE (product no: RPN2232). Heparin sodium salt was obtained from Stemgent (cat no: 07980). BL21(DE3) competent cells were purchased from Life Technologies (cat no: C6000-03). IPTG was purchased from Invitrogen (cat no: AM9464). NI-NTA beads were purchased from Qiagen (cat no: 1018611). Pierce BCA assay kit was purchased from Thermoscientific (23225). Acrodisc 0.8uM syringe filters were purchased from PALL Life Sciences (cat. no. PN4618), Protease inhibitor cocktail set VII was purchased from Calbiochem (cat no. 539138-1SET), 3000 Da MWCO centrifugal spin filters were purchased from Amicon (cat no. UFC900324). RPMI 1640 was purchased from Life Technologies (cat no. 72400-047), FBS from Life Technologies (cat no. 10082-147), Pen/Strep from Life Technologies (cat no. 15140-122), and IL-3 from R&D. 1mL HiTrap Heparin HP columns were purchased from GE (cat no. 17-0406-01).

Site-directed mutagenesis

FGF-2 constructs were cloned and purified as previously described [123]. Site directed mutagenesis was performed using Agilent Quik change II XL kit (Agilent cat no. 200522-5). The primer list is included in Table 2.1. The introduction of the mutations in the bFGF construct was confirmed through sequencing with Genewiz T7 primer 5'-d(TAATACGACTCACTATAGGG)-3'.

bFGF (aa #)	Primer (5'--> 3)'
Y132A	gaacgactcgagtctaataacgccaatacttaccggtcaaggaa
Y132A_anti	ttccttgaccggtaagtattggcgttattagactcgagtcgttc
Y53A	accccaagcggctggcctgcaagaacgggg
Y53A_anti	ccccgttcttgcaggccagccgcttggggt
R138A	taactacaatacttaccggtcagcgaaatacaccagttggtatgtg
R138A_anti	cacataccaactggtgtatttcgctgaccggtaagtattgtagtta
K164A	ccaaaacaggacctgggcaggcagctatactttttcttccaa
K164A_anti	ttggaagaaaaagtatagctgcctgcccaggtcctgttttgg
K154A	tggcactgaaacgaactgggcagtatgcacttggatccaaaa
K154A_anti	ttttggatccaagtgcatactgcccagttcgtttcagtgcca
F124A	ctaaatgtgttacagacgagtgtttctttgctgaacgactc
F124A_anti	gagtcgttcagcaaagaaacactcgtctgtaacacatttag
Y53F	cccgttcttgcagaacagccgcttggg
Y53F_Anti	cccaagcggctgttctgcaagaacggg
F124Y	agactcgagtcgttcataaaagaaacactcgtctgtaac
F124Y_Anti	gttacagacgagtgtttcttttatgaacgactcgagtct

F124E	attagactcgagtcgttcctcaaagaaacactcgtctgtaacacatttagaagc
F124E_Anti	gcttctaaatgtgttacagacgagtgtttctttgaggaacgactcgagtctaat
F124D	gactcgagtcgttcatcaaagaaacactcgtctgtaacacatttag
F124D_Anti	ctaaatgtgttacagacgagtgtttctttgatgaacgactcgagtc
D119L	ctcgagtcgttcaaaaaagaaacactctaatgtaacacatttagaagctagtaatct
D119L_Anti	agattactagcttctaaatgtgttacattagagtgtttcttttttgaacgactcgag
D119K	gagtcgttcaaaaaagaaacactcctttgtaacacatttagaagctagtaa
D119K_Anti	ttactagcttctaaatgtgttacaaaggagtgtttcttttttgaacgactc
F122A	tagactcgagtcgttcaaaaaaggcacactcgtctgtaacacatttag
F122A_Anti	ctaaatgtgttacagacgagtgtgccttttttgaacgactcgagtcta
F122Y	tattagactcgagtcgttcaaaaaaataacactcgtctgtaacacatttaga
F122Y_Anti	tctaaatgtgttacagacgagtgttatttttttgaacgactcgagtctaata
F122E	gttattagactcgagtcgttcaaaaaactcacactcgtctgtaacacatttagaagc
F122E_Anti	gcttctaaatgtgttacagacgagtgtgagttttttgaacgactcgagtctaataac
K55A	gaagaagcccccgttcgcgcagtacagccgcttg
K55A_Anti	caagcggctgtactgcgcgaacgggggcttcttc
R149A	aagtttatactgcccagttgctttcagtgccacataccaac
R149A_Anti	gttggtatgtggcactgaaagcaactgggcagtataaactt
K158A	tttctgcccaggtcctgttgcggatccaagtttatactgc
K158A_Anti	gcagtataaacttggatccgcaacaggacctgggcagaaa

Table 2.1 FGF mutagenesis primers

Expression, isolation and purification of FGF in *E.coli*

25mL Luria broth (LB) cultures, containing ampicillin, were inoculated with BL21(DE3) competent cells expressing wild type and mutant FGF-2 genes were grown overnight at 37°C. The following day the 25mL starter culture was inoculated into 1L of LB (plus ampicillin) and grown to an OD_{600} of 0.6-0.8. Once the OD reached expression, the construct was induced with 1mM iso-pentyl-thio-galactoside (IPTG) for 4 hours and harvested. The cell pellet was resuspended in 20 mL of Ni-NTA purification wash buffer (20 mM Tris, 500 mM NaCl, 5 mM imidazole, pH 8.0). The resuspended bacterial culture was placed in an ice bath and sonicated using a Qsonica sonicator (model no.

Q700, power 700 W) at an amplitude of 50 with pulse at 20s and rest at 5s. This was repeated for a total process time of 3 min. After sonication the sample was centrifuged at 15,000 xg for 30 mins to remove insoluble material. The supernatant was filtered through a 0.8uM syringe filter and spiked with 250uL of the CalBiochem protease inhibitor cocktail set VII. This sample was incubated with Ni-NTA beads (previously washed in water and equilibrated in Ni-NTA purification wash buffer) and left rotating end-over-end at 4 degrees Celsius overnight. The Ni-NTA beads were transferred to spin columns, washed 3 times in wash buffers and eluted sequentially in various mixtures of Wash and Elution buffer (20 mM Tris, 500 mM NaCl, 200 mM imidazole, pH 8.0), i.e. 10%, 20%, 40%, 80% and 100% elution buffer. All the elution fractions were collected. The 40%, 80%, and 100% fractions were pooled into 15 mL samples and buffer exchanged into 20mM Tris, 50mM NaCl pH 7.4 using 3000 Da MWCO centrifugal spin filters. A Bicinchoninic Acid (BCA) assay was performed to estimate protein concentration after purification. The sample was stored at 4°C for up to two weeks.

Cell culture

Ba/F3 cells expressing the human FGFR1c were maintained as a suspension culture in RPMI 1640 containing 10% fetal bovine serum, 10ng/mL IL-3 or 10% WEHI-3 conditioned media, and penicillin/streptomycin. Cells were maintained at a density between $1x10^5$-$1x10^6$/mL. WEHI-3 Conditioned media was generated as described in Padera et al. 1999 [123]. All cells were cultured in a 37°C / 5% CO_2 incubator.

Proliferation assays

BaF3/FGFR1c (F32) cells were harvested, washed twice in culture media lacking IL-3, and counted. Cells were then seeded in IL-3 deficient media at a density of $1x10^5$ cells/mL in a 24 well plate. Various doses (0.01nM-100nM) of FGF-2 or FGF-2 mutants were incubated with F32 cells, and allowed to incubate for 72 hours. In cases where heparin was added a dose of 500 ng/mL was used. After 72 hours, the cell number was counted using a Beckman Coulter Multisizer 4 particle counter (only particles w/ a volume >40um^3 were counted). Cell number (x) was normalized between 0 and 1 ((x-min)/(min-max)) to yield the value for the proliferation index.

Heparin Affinity Chromatography

To measure heparin affinity of FGF-2 and the mutants, we employed a heparin affinity chromatography FPLC assay. The 1mL HiTrap Heparin HP column was equilibrated in Wash buffer (10mM Sodium Phosphate, 100mM NaCl pH 7.4). Next, 200ug of purified protein was loaded onto the column at a flow rate of 1mg/mL. The column was washed briefly with 2 column volumes of wash buffer, and then subjected to a linear salt gradient ranging from 0.1 -2M NaCl. The linear gradient was formed by mixing the wash buffer with an Elution buffer (10mM Sodium Phosphate, 2M NaCl pH 7.4). Twenty-four 0.5mL elution fractions were collected and from those fractions, protein identity was confirmed by SDS-PAGE/western blot using antibodies against the 6xHis Tag fused to FGF-2 construct.

CD Spectroscopy

CD Analysis was performed using an Aviv Model 202 Circular Dichroism spectrometer. A standard wavelength read, between 195nm-260nm at 25°C was performed in a Hellma Cell CD cuvette (1mm pathlength; part number 111- QS). The sample diluted in 20mM Tris pH7.4 and was read at 300ug/mL (unless otherwise noted). The background was subtracted using the blank buffer read and the final data expressed in Molar Ellipticity ($[]_M$x10^{-3} (degcm^2/dmol^{-1})).

Significant Interaction Network Calculations

Coordinates for each point mutation were obtained using the automodel class of Modeller v9.10. The sequence and coordinate structure of FGF2 (pdb 4FGF) were used as a template in homology modeling to produce coordinate files for each point mutation studied. Further, the solvent accessibility of the residues was calculated using DSSP server (http://www.cmbi.ru.nl/dssp.html). For each residue the inter-residue interactions was calculated by incorporating putative hydrogen bonds (including water-bridged ones), disulfide bonds, pi-bonds, polar interactions, salt bridges, and Van der Waals interactions (non-hydrogen) occurring between pairs of residues within a threshold distance was computed as described previously [129]. These data were assembled into an array of eight atomic interaction matrices. A weighted sum of the eight atomic interaction matrices were then computed to produce a single matrix that accounts for the strength of atomic interaction between residue pairs, using weights derived from relative atomic interaction energies [129]. The inter-residue interaction network calculated in this fashion generates a matrix that describes all the contacts made by a given residue with spatial proximal neighboring residues in their environment. Each element *i, j* is the sum of the path scores of all paths between residues *i* and *j*. The degree of networking score for each residue was computed by summing across the rows of the matrix, which was meant to correspond to the extent of "networking" for each residue. This interactional relationship is represented using a two-dimensional network diagram. The degree of networking score was normalized with the maximum score for each protein so that the scores varied from 0 (absence of any network) to 1 (most networked).

Results

Model system selection

In order to investigate if our SIN tool could be used to identify functionally important residues in a heparin binding site, a well-studied model system was required. The fibroblast growth factor family (FGF) of heparin binding proteins was chosen for our studies. Specifically, the work here was conducted using FGF-2 or bFGF. This protein was selected on the basis of three criteria: i) crystal structures for bFGF in complex with heparin and its receptor exist, ii) the heparin binding specificity is well-characterized, and iii) heparin-bFGF has been studied structurally and biochemically, so several key residues are known. Furthermore, bFGF has been extensively studied and functional residues critical for FGF dimerization and receptor binding are well characterized. The next two sections will cover a brief, but relevant, background of the FGF family and bFGF.

FGF family and biological function

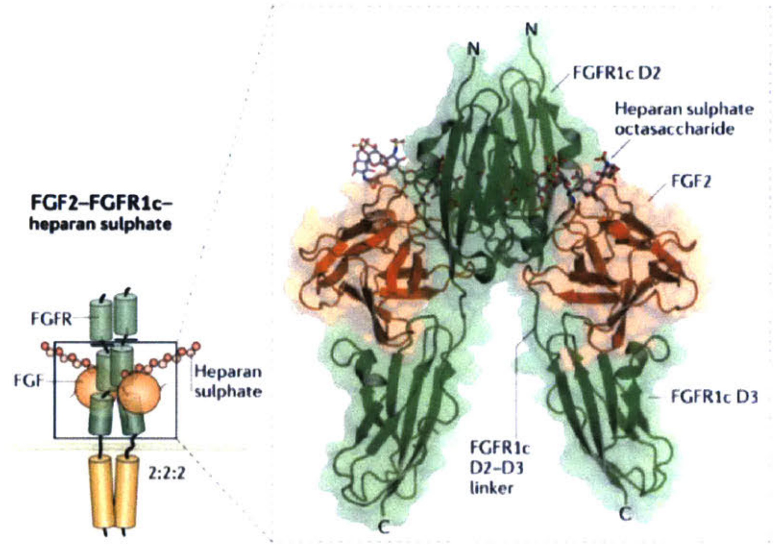


Figure 2.1 **Structure of the FGF ternary complex**. Cartoon schematic and crystal structure (PDB ID: 1FQ9) rendering of the active ternary signaling complex, composed of FGF2-HSGAG-FGFR1c in a 2:2:2 ratio. In this interaction, each FGF-2 ligand directly interacts with FGFR1's DII and DIII domains. Furthermore, FGF and the FGF receptor both have extensive direct interactions with HSGAGs. This figure is adapted from [131]

FGF is a family of growth factors, comprised of 18 members (FGF1-FGF10 and FGF16-FGF23) and they can be grouped into 6 subfamilies based on phylogeny and sequence homology. The 6 subfamilies are FGF1 (FGF1 & -2), FGF4 (FGF4, -5,-6), FGF7 (FGF3, -7, -10, -22), FGF8 (FGF8, -17, -18), FGF9 (FGF9, -16, -20) and FGF19 (FGF19, -21, -23) [122,131]. FGFs are known to regulate many diverse biological processes such as embryonic development, angiogenesis, wound healing, tissue regeneration, metabolic signaling, inflammation and cancer [122]. The action of FGFs are exerted through their interaction with the FGF receptors (FGFR1,-2,-3,-4), a family of tyrosine kinases [122,131,132]. Importantly, these interactions are dependent on Heparan sulfate glycosaminoglycans (HSGAGs). The active signaling complex, also known as the ternary complex, is comprised of dimerized FGF bound to HSGAG, which are both bound to a dimerized FGFR in a 2:2:2 ratio (Figure 2.1) [87]. Almost all the members of the FGF family bind HSGAGs with high affinity, although they differ in their affinity and specificity for various HSGAG structures. Structurally, FGFs are comprised of a homologous core region consisting of ~125 amino acids, ordered into 12 anti-parallel β-strands ultimately yielding a β-trefoil core structure [122]. Despite the structural homology, FGF family members interact with various isoforms of the FGF receptors and have unique heparin binding specificity/affinity. Understanding the structural determinants which underlay FGF-heparin interactions, FGF-FGFR interactions, and FGF dimerization is of interest, as elucidating structural mechanism of FGF signaling and regulation could yield important insight into FGF biology.

Structure and function of FGF-2

FGF-2 or basic FGF (bFGF), is a member of the FGF family that has been extensively characterized. bFGF can interact with several FGFR isoforms (i.e. FGFR1c, 2c, 3c, 1b) [132]. When in complex with the FGFR, FGF makes direct interactions with the FGFR's D-II and D-III domains (Figure 2.1). On bFGF these domains are termed D-II and D-III binding domains, and have been characterized previously [133]. Importantly, FGF-2 interacts directly with HSGAGs with high affinity and it also binds HSGAG in the context of the active ternary signaling complex [87]. The specificity and structural determinants of the high affinity interaction between HSGAGs and FGF-2 have been studied at multiple levels, including: the affinity for various HS with different structures, identifying the amino acids critical for HSGAG binding, and the conformational requirements for heparin to bind FGF-2 [91,92,123,133].

It is worth noting that uncovering the structural determinants critical for HSGAG-protein interactions is not trivial, and our understanding of these interactions has evolved over the last two decades. For many years, it was generally assumed that interactions between FGFs and HSGAGs was purely driven by the electrostatics. Indeed, HSGAGs are highly negatively-charged and the heparin-binding sites on protein are often characterized by patches of positively-charged basic residues spatially oriented to make ionic contacts with sulfo or carboxyl groups on HSGAG chains [134]. In fact, a few consensus sequences of heparin binding proteins have been identified (XBBXBX, and XBBBXXBX, where X represents any amino acid and B are basic residues) [134]. This rudimentary view of heparin-protein interactions was challenged when it was identified that anti-thrombin, a heparin-binding protein (HBP), bound with high affinity to specific HSGAG sequences. These results suggested that HBPs can discriminate between different HSGAGs and that the fine structure (or "sequence") of HSGAGs play a critical role in regulating HBPs biological function [90]. Consequently, the field now appreciates that high affinity interactions are achieved through ionic interactions, as well as additional contacts between the protein and HSGAG, such as hydrogen bonding and van der Waals contacts [92]. The latter confirms that the structural determinants of HBP-HSGAG interactions are more nuanced than negatively charged polymers interacting with a patch of positively charged residues.

Previously, our lab had studied the FGF-2-HSGAG interaction. Specifically, our lab determined the conformational requirements for FGF-2 to bind with high affinity to heparin [110]. Using a crystal structure of FGF-2 in complex with a hexasaccharide comprised of a trisulfated dissacharide repeat of $[I_{2S}H_{NS,62}]_n$ (I= α-L-iduronic acid; H=α-D-glucosamine), conformational analysis showed that the FGF-2 bound HSGAG contained a "kink" spanning a trisaccharide H-I-H motif. This kink in FGF-2 bound HSGAG, allows for optimal ionic interaction between basic residues in the heparin binding pocket of FGF-2, as well as contacts with residues on the periphery of the binding pocket. Additionally, optimal van der Waals interactions were achieved in this conformation. Importantly, kink formation in HSGAGs is thought to be a key structural attribute of FGF bound HSGAG, and is hypothesized to be conserved amongst all FGF-HSGAG interactions [91,110]. In this model, the unique specificity and affinity of each HSGAG interaction is thought to be regulated by protein surface topology and the unique distribution of basic residues in the heparin binding site of FGFs.

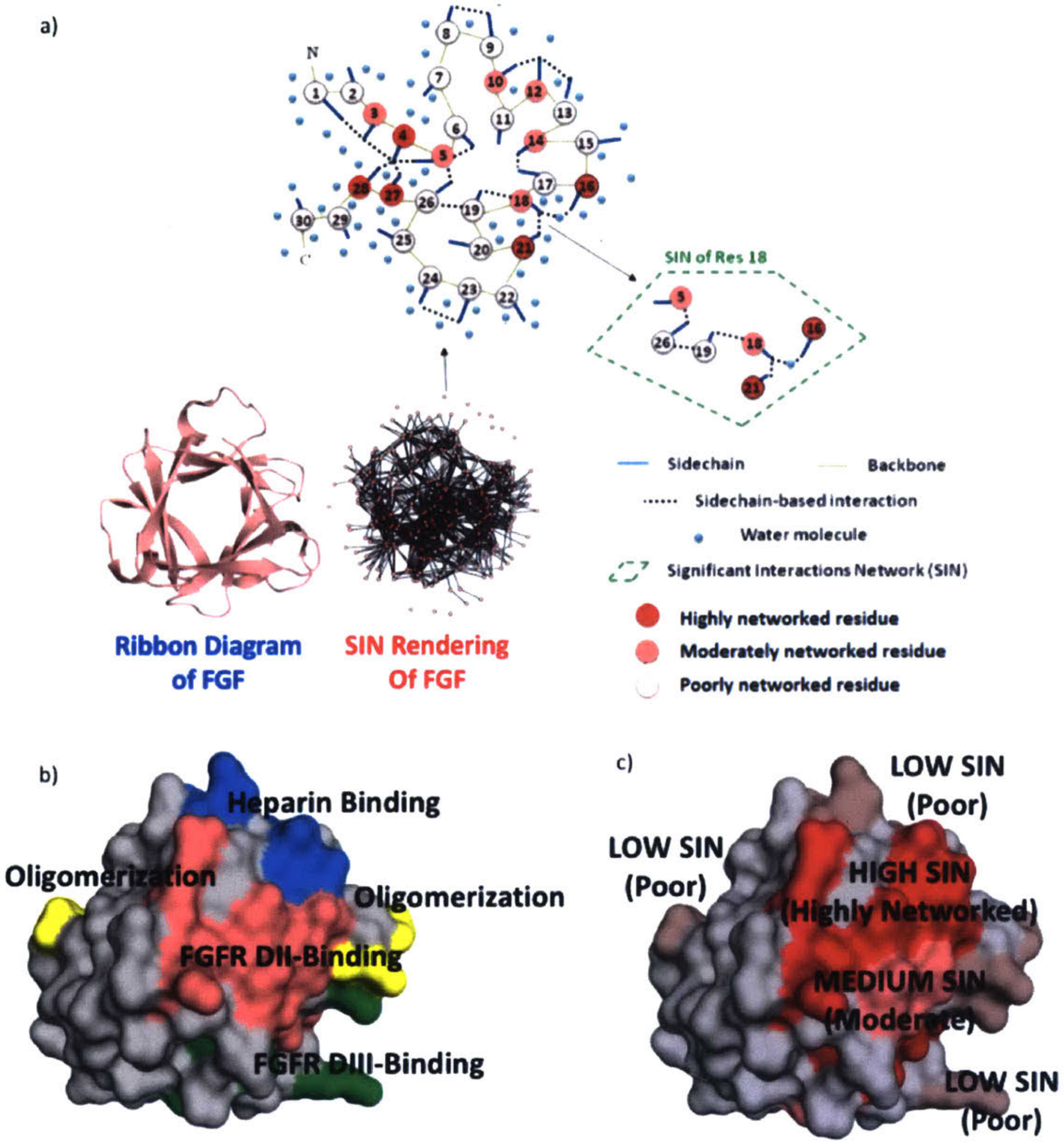


Figure 2.2 **SIN analysis of bFGF**. a) Depicted here is crystal structure of bFGF (PDB ID-4FGF) in a Ribbon diagram. Note the β-trefoil core structure of bFGF. Adjacently, the SIN rendering and network diagram of bFGF is shown. An enlarged image of the results of the network analysis is pictured above. In this diagram each node is residue, and each edge is an interaction (delineated in the legend). For each residue dark red= high score, faint red= low score. b) Surface rendering of bFGF. Amino acids are colored according to regions that are known to have some function red= FGFR-DII binding residues, yellow=residues involved in oligomerization, green= FGFR DIII binding residues, blue= heparin binding residues. c) Amino acids are colored according to inter-residue network score, Dark red= high score, faint red= low score.

SIN analysis of FGF-2

The goal of this study is to investigate if our inter-residue network analysis tool, SIN, could aid in the identification of key residues in the heparin binding site of bFGF. Residue interaction network scores for each residue in FGF were calculated using the crystal structure (PDB ID= 4FGF). The SIN score takes both inter-residue contacts and solvent accessibility into account (see experimental methods for details). Residues with a high SIN score have a high degree of network connectivity. There are several interpretations of the scoring. From previous work, it is known that high scoring residues are structurally constrained to mutate [129]. From an evolutionary perspective this may mean that these residues play a critical role in a protein's structure or are critical for protein function. In the case of bFGF, a SIN score threshold was defined; above that threshold low, medium, and high scoring residues were identified. As depicted in Figure 2.2a and Figure 2.2c, these residues were colored using a red scale based on their low-, medium-, or high SIN scores. Amino acids in the D-II binding site on FGFs were found to have a medium to high SIN score, whereas residues in the D-III binding domain and the oligomerization domain had a low SIN score. Interestingly, the heparin binding site contains low, moderate, and high scoring residues. Finally, three additional residues with a high SIN score were identified (D90, F93, and F95) which had no previously known significance.

The SIN score for other members of the FGF family with crystal structures (FGF-7, -8, -9, -10, -19) were also calculated (Figure 2.3). While there were some subtle differences, the aforementioned trend in SIN score largely held true; the D-II binding domains were highly networked, the D-III domains were lowly networked, and the heparin binding site contained low, moderate, and highly networked residues. These results agree with what has previously been reported about the FGF family. The D-II domain, also known as the primary receptor binding site, is the high affinity FGFR binding site. Several residues comprising the D-II binding domain (Y24, E96, N101, Y103, and L140) are known to be conserved amongst all the FGFs [133]. It is, thus, not surprising that these residues, as well as others comprising the D-II binding site, have a high network score. On the other hand, the D-III binding site (or secondary receptor binding site) is not conserved in a sequence space, but appears to share similar network properties across all FGFs, where the D-III residues have low to moderate network score. These results make sense, as each FGF subfamily is known to bind to different FGFR receptors [132]. The FGF receptors often contain splice variants of the D-III domain. Consequently, *a priori* identification of the D-III residues from sequence analysis is not possible due to the sequence diversity present in FGF D-III domain. A network view, however, appears to identify some of these functional residues in the D-III binding site on all FGF family members. Lastly, in the putative heparin binding region of each of the FGFs studied, the residues which are thought to anchor the H-I-H kink motif were found to be low to moderately networked.

The fact that the D-II, D-III, and heparin binding regions are detected using SIN analysis suggests that a network view of the FGF family captures key protein regions that are structurally conserved. It is worth noting that there is only ~10-55% sequence identity between all the 18 FGF family members, so it is challenging to identify functional residues from a sequence standpoint [133]. These results highlight that our SIN analysis provides complementary information to a bioinformatics guided sequence analysis.

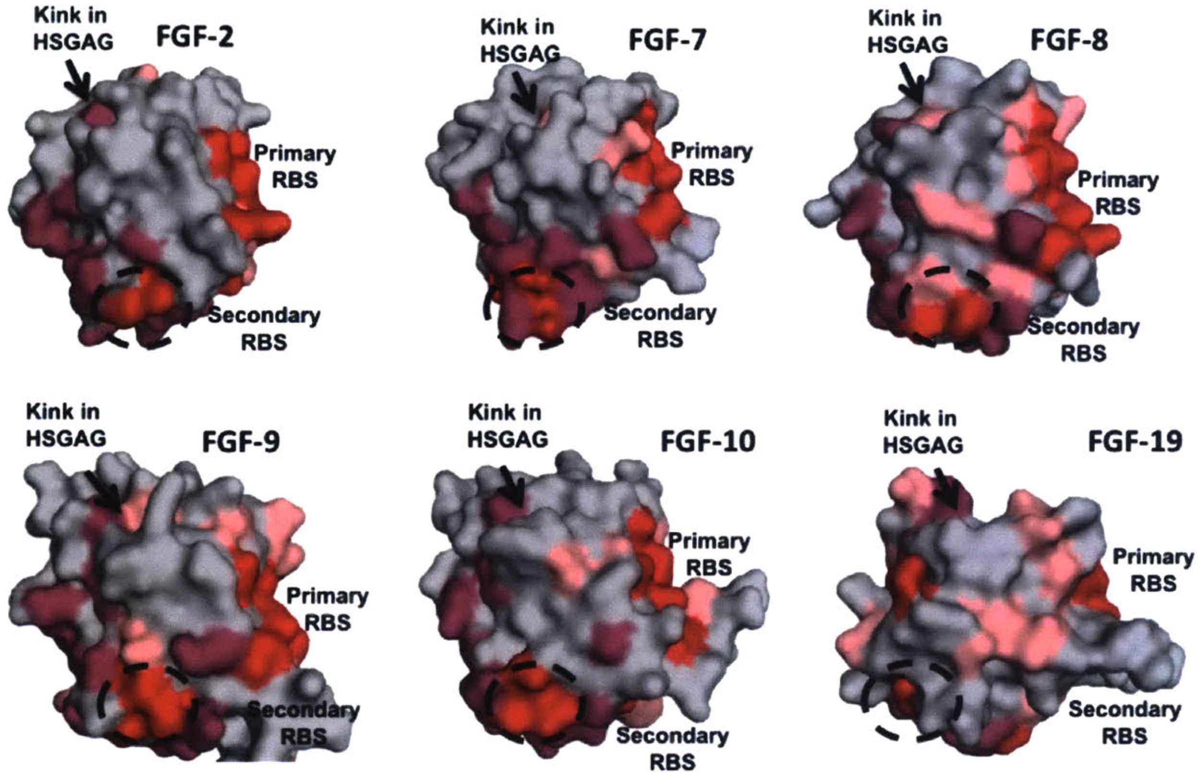


Figure 2.3 **Surface rendering of FGF family members (FGF-2, FGF-7, FGF-8, FGF-9, FGF-10, and FGF-19).** Amino acids are colored according to inter-residue network score, Dark red= high score, faint red= low score. Note that in all cases the primary RBS contains moderate to highly networked regions, and the secondary RBS contains low-moderately network regions. The heparin binding region, including the putative "kink" interacting region is highlighted using an arrow and is composed of low-moderate network scores. Finally, the region in the dotted circle is a highly networked region which has no known function.

SIN-guided mutagenesis studies

To confirm that SIN analysis can identify residues that are critical to protein function, mutations at select FGF residue positions were made and tested. Several moderate to high SIN scoring residues were selected from each of the representative functional areas, including the D-II, D-III, heparin binding regions, and a high SIN scoring region with no known function. In order to select the residues and specific mutations a homology model, containing the putative mutation, was constructed and a SIN score for the mutated residue was re-calculated. Mutations that had a significant impact on the network score (where there was a >5 fold decrease in the mutant SIN score change) were introduced into FGF-2 and expressed recombinantly in *E. coli.*

These mutations include: Y24A, Y103A (D-II region), R109A (D-III region), K125A, K135A (heparin binding region), and F95A (unknown region).

Each mutant construct was created and expressed in *E.coli,* purified using nickel chromatography, and subjected to various quality control and functional tests. Briefly, the identity of the FGF-2 mutants were tested using western blot and MALDI-MS (data not shown). Protein expression and circular dichroism (CD) were used to assess impact on protein stability (Figure 2.4a and c). Heparin binding properties were tested using heparin affinity chromatography (Figure 2.4d) and biological activity was tested using a cell proliferation assay (Figure 2.4b). The biological activity assay measures the mitogenic ability of each FGF mutant using a cell line, BaF3, that expresses FGFR1c [123]. The results of these analyses are represented in summary Table 2.2.

All of the mutants, with the exception of F95A (discussed later), expressed well and gave a CD spectra profile consistent with WT bFGF. FGFR-binding mutants (Y24A, Y103A, R109A), were first tested for their heparin binding properties. Y103A (D-II mutant), and R109A (D-III) showed no impact on FGF-heparin interaction, in concordance with previously published literature. Y24A (which was selected as a D-II mutant) showed a significant impact on heparin binding. It is unclear how or why this occurs. Y24 is a residue that is known to be conserved in almost all FGFs, yet it has never been reported as important structural determinant for heparin binding. It is possible that Y24A makes critical inter-residue interactions with other keys structural determinants in the heparin binding site, such as K26. An alternative explanation is that Y24 plays a critical role in structural stability or folding in FGF. However, this explanation is unlikely as Y24A showed a normal CD spectra and there was no effect on the expression rate. In biological activity assays, both D-II mutants showed ~10-14 fold increase in EC_{50}. These results are consistent with those previously reported [123]. Conversely, the R109A mutant showed no change on biological activity, which is consistent with its known function as a residue making contacts with the D-III domain of FGFR.

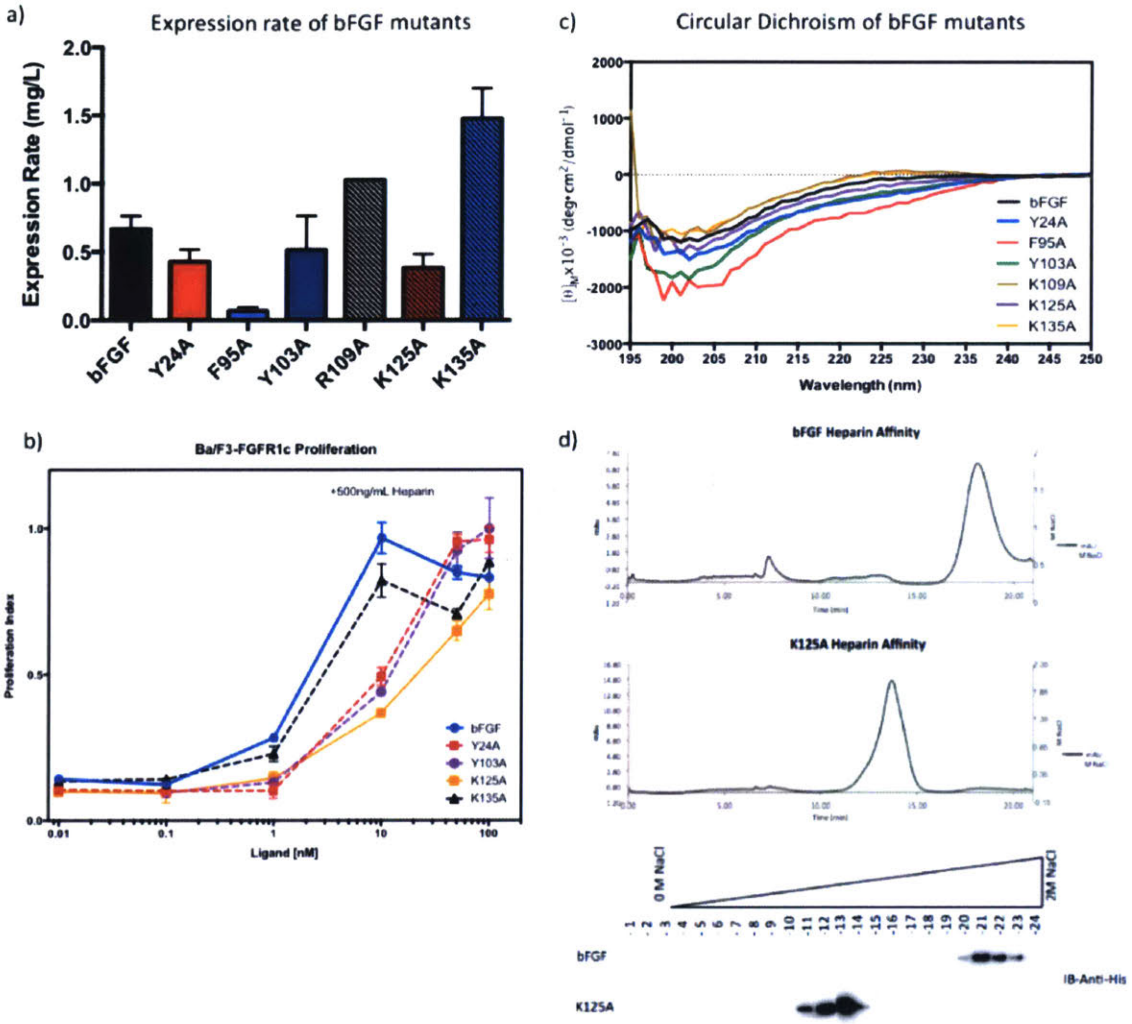


Figure 2.4 **Integrated experimental analysis to study bFGF** a) Expression rates of bFGF and mutant constructs. Bar graph shows the protein expression yields after purification. Error bars represent standard deviation from the mean. b) Cell proliferation assay with BaF3-FGFR1c cells in the presence of 500 ng/mL of heparin. Proliferation index of each mutant constructs is expressed a function of ligand concentration is shown in the colored lines. Error bars represent standard deviation from the mean. c) Circular Dichroism Spectra of bFGF and mutant constructs. Molar elipticity of the protein is shown as a function of wavelength for each sample. d) representative FPLC Heparin affinity chromatogram of WT bFGF and the K125A mutant. Thin dotted line represents molar concentration of salt added to the column (Right y-axis). Thick solid line represents amount of detected protein A_{280} eluted off the heparin column (Left y-axis). Below are western blot images of elution fractions from bFGF and K125A mutant. Linear salt gradient from 0-2M illustrated by triangle with corresponding numbered elution fractions below in which protein was collected.

It is worth noting that one limitation with our experimental framework is that a biological activity assay is used as a proxy for FGF receptor binding and activation. There is a known interplay between FGF and heparin where either molecule can functionally compensate for the other when inducing receptor dimerization and mitogenesis [123]. For instance, it is not possible to know if the biological impact of the Y24A mutation was driven by changes in the FGF-FGFR interaction, FGF-heparin interaction or both. A biochemical assay that directly measured FGF-FGFR1c interaction would have provided additional clarity on this matter. However, the focus of this study was to test the ability of SIN to identify key functional residues in a model protein system and was not focused on uncovering FGF biology.

The two heparin binding mutants, K125A and K135A, were also subjected to biological activity assay and heparin binding assay. Both K125A and K135A decreased heparin binding. K125A had a much larger decrease in heparin binding relative to K135A. In biological activity assays, K135A did not have an impact on EC_{50}, whereas K125A had a 30-fold increase in EC_{50}. Interestingly, K125A had the largest impact on biological activity of any residue tested. As an aside, the two residues with the largest fold change in SIN score from WT (Y103A and K125A) had the largest change in biological activity, suggesting the SIN fold change may be a useful metric to predict or prioritize which residues are functionally critical. Further validation would be required to prove the latter.

The results of the heparin binding mutants yield important and novel insights into the structural requirements for high affinity heparin-FGF interactions. In earlier work, all heparin binding residues were considered functionally equivalent, but the results here suggest that certain residue contacts are functionally more important than others. These can be reasoned structurally, as K125 sits deep in the pocket of the heparin binding domain and is known to be a key residue responsible for anchoring the kink motif H-I-H of the HSGAG. Specifically, it is known to make key contacts with 2-O-sulfate groups and N-sulfate groups of the H-I-H trisaccharide, while the K135A sits on the periphery of the binding pocket making contact with other sulfo- and carboxy-groups on the HSGAG chain. Previous work in our lab showed that the kink adopted by the HSGAG was a crucial structural determinant for creating high affinity FGF-Heparin interactions. The results here suggest that the corresponding residues on FGF responsible for facilitating/anchoring the "kinked" HSGAGs are also critical for high affinity FGF2-heparin interactions. Such nuances regarding the protein contribution to kink anchoring have never been appreciated prior to this analysis.

Finally, F95 is a high SIN scoring residue, and the F95A mutation had a large impact on the SIN score. This mutation appears to play an important role in the structural stability of the bFGF, as the mutation significantly decreases the expression. Furthermore, the protein expressed appears to suffer structural deficits, as the CD spectra appears to deviate significantly from the WT protein (Figure 2.4a and Figure 2.4c). From these analyses, it is plausible that the F95 position plays a pivotal role in the structural stability or folding of bFGF, however, further analysis is required to ultimately draw a conclusion about the function of the unknown region.

Construct	Functional region	WT SIN	Mutant SIN	Expression Rate	Circular Dichroism	EC_{50} (nM)	Heparin Affinity
bFGF (WT)	N/A	N/A	N/A	-	N/A	1.21	-
Y24A	D-II	0.536	0.102	-	N.C.	10.6	↓↓
Y103A	D-II	0.59	0.037	-	Slight deviation from WT	14.89	-
R109A	D-III	0.507	0.091	↑	N.C.	1.6	-
K135A	Heparin-Binding	0.259	0.046	↑↑	N.C.	1.33	↓
K125A	Heparin-Binding	0.114	0.008	-	N.C.	31.51	↓↓↓
F95A	Unknown region	0.568	0.074	↓↓	Large deviation from WT	NT	NT

Table 2.2. **Summary table of the mutants identified from SIN guided mutagenesis.** (-) indicates no change relative to WT. (↑ or ↓) indicates an increase or decrease relative to WT. The number of arrows indicates the magnitude of the change. N/A = not applicable, NT= not tested, and N.C.= no change relative to WT. EC_{50} values show the concentration of ligand needed to induce half the maximum of proliferation. Values were obtained by performing a four-parameter fit on the proliferation index data for each construct. All values have R^2 values greater than or equal to 0.96.

Conclusions

The objective of this study was to determine if SIN analysis could be used to detect key functional residues in a protein, with the goal of integrating this tool into future studies of glycan-GBP interactions. SIN analysis identified low, medium and highly networked residues on bFGF and these residues corresponded to areas with known protein functions such as FGFR binding and heparin binding. Furthermore, SIN analysis of other members of the FGF family, showed that key functional residues in each FGF's analogous D-II, D-III, and heparin binding were also identified. This suggests that SIN, can capture structurally conserved region despite low or minimal homology in a sequence space. Thus, SIN provides orthogonal information to a bioinformatics or sequence analysis based approaches to investigate evolutionary conservation.

Using SIN-guided mutagenesis, several putative mutations were identified and created. The mutations selected validate that SIN identifies residues that are functionally important as the mutations that were created impacted bFGF's function in heparin binding, biological activity, or structural stability. These results suggest that SIN is a tool that could be used for efficient, *a priori* identification of key functional residues with nothing more than a crystal structure or homology model.

Though it was not the explicit purpose of this study, the SIN guided mutagenesis did yield some interesting biological insights and observations that warrant further investigation. First, given that higher SIN scoring residues correlated well with known functional regions, it is surprising that a patch of high SIN scoring residues was identified with no known biological function. Our preliminary results suggest that this region might play an important role in the structural stability of FGF, however further investigation is required to make this conclusion. Additionally, new insight into which residues on bFGF are most critical for the high affinity heparin interaction was uncovered. Specifically, we uncovered K125 as a key structural determinant of the bFGF-heparin interaction, and ultimately, its biological function.

In addition to the ability of SIN to identify functional residues and structurally conserved regions studied here, previous work has shown that SIN identified amino acids that are structurally constrained to mutate as well as long range interactions between residues. Taken together, the network analysis provided by SIN enables new, useful insights and should be incorporated into the integrated experimental framework used to study the structural determinants of GBP-glycan interactions.

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Rahul Raman, Sylvia Atsaves, David Johnson, Ashley Santelises, Prithwis Mukhopadyay and Ram Sasisekharan for their contributions to this work. This work was funded in part by the National Institutes of Health (T32 ES007020-38) and MIT UROP Funding Office and the Maurice N. Katz Family Fund for UROP

Chapter 3 : Structural determinants for naturally evolving H5N1 hemagglutinin to switch its receptor specificity

Summary and Significance

As mentioned previously, the study of GBPs is complex. Consequently, understanding the structural determinants that govern the specificity and affinity of glycan-GBP interactions requires a unique, integrated experimental approach. In chapter 2, I implemented an integrated experimental analysis comprised of structural analysis and biochemical and functional assays with the goal of studying FGF-heparin interactions. I incorporated an inter-residue network analysis tool (SIN) into this integrated framework and demonstrated the utility of this tool. Specifically, I showed its ability to quickly identify key residues for glycan-protein interactions and yield novel insights into the structural determinants of the bFGF-heparin interaction. Leveraging this improved experimental framework and lesson learned in chapter 2, I applied our approach (including the incorporation of SIN analysis) to study the structural determinants for avian-adapted H5N1 hemagglutinins to switch their receptor specificity and bind with high affinity to human glycan receptors.

Highly pathogenic H5N1 continues to pose a severe pandemic threat to humans. Of all the avian viral subtypes highly pathogenic H5N1 is of upmost concern due to the high case fatality rate (~60%) associated with H5N1 outbreaks. Importantly, H5N1 has never adapted to human hosts and efficient respiratory droplet transmission has never been observed. Pandemic viruses often originate from avian hosts, where the virus has gained adaptations that enable it to infect, replicate and transmit in humans. Hemagglutinin (HA) is a GBP which regulates the first step of IAV binding and entry into the cell. Adaptation of the IAV HA protein is a critical step regulating the host switch. It is critical for governing host tropism, where avian-adapted HAs bind with high affinity to $\alpha2\rightarrow3$ sialylated glycans (or avian receptors) capable of adopting a cone-like topology, and human adapted HAs bind with high affinity to $\alpha2\rightarrow6$ sialylated glycan receptors (or human receptors) capable of adapting an umbrella like topology. Furthermore, for HAs, a loss of avian receptor binding in conjunction with a gain in high affinity binding to long α2→6-linked human glycan receptors is known as a quantitative switch and is correlated with efficient respiratory droplet transmission. Importantly, the structural determinants required for H5N1 HA to undergo this quantitative switch remain unknown. Elucidating the structural requirements for human adaptation of H5N1 is critical for pandemic preparedness, as uncovering these determinants could help IAV surveillance or enable creation of H5N1 vaccines and therapeutics.

In chapter 3, using an integrated approach comprised of structural modeling, bioinformatics, inter-residue network analysis (SIN), quantitative glycan-binding assays and tissue staining assay, I investigate the structural determinants for currently evolving H5N1 HAs to quantitatively switch their receptor specificity from avian to human. First, using a structural comparison between avian adapted HA and its nearest human adapted phylogenetic neighbor (H2), four structural features required for the human adaptation of H5N1 HA were identified. Next, using network analysis (SIN) to investigate the local network of residues comprising the structural features, potential mutational paths for acquisition of the human adapted HA features were identified. The aforementioned analysis took the structural features and mapped them in a network space, where each feature was reduced to a set of network contextualized mutations required for human-adapted feature acquisition. Using bioinformatics analysis, the structural features needed for human adaptation were assessed in the context of recently identified and currently evolving H5N1s. Several strains from representative clades isolated from recent H5N1 outbreaks (2007-2010), were found to naturally contain one or more of the human-adapted features. These HAs were selected, expressed recombinantly and the quantitative affinity for avian and human glycan receptors was measured. Next, in order to validate that the structural features for the human adaptation of H5 could induce a quantitative switch, mutations that added one or more structural features were introduced into currently evolving H5 HA. Indeed, introduction of structural features in the several naturally evolving H5 HAs could induce a quantitative switch. In one H5, dkEgy10 (isolated in 2010), a single amino acid mutation was sufficient to induce a quantitative switch suggesting that certain H5N1s

HAs are close to human adaptation. This mutant showed extensive binding to the apical surface of the upper respiratory tract, consistent with the tissue binding signature of other human-adapted HAs.

The goal of this study was to apply an integrated analysis to study the structural determinants for H5N1 to switch its receptor specificity from avian to human. Indeed, four structural features required for human adaptation of H5 were identified and validated. Interestingly, some of these features have been acquired naturally in recently circulating H5N1 strains and some strains are especially close to human adaptation. In these cases, as few as one mutation could confer a quantitative switch in binding preference from avian to human glycan receptor. It is worth noting that prior to this study, there were two independent studies showing efficient respiratory droplet transmission of laboratory adapted H5N1s. These studies identified several "hallmark mutations" required for human adaptation and suggested monitoring H5s for acquisition of these mutations as a viable surveillance strategy. Both of these studies used old H5s (i.e. 2004 and 2005 isolates) and there has since been significant sequence evolution in H5 HAs. This rudimentary approach to surveillance is not robust, as it ignores the context of currently evolving H5 HAs. In fact, introduction of several key hallmark mutations into more recent H5s was not able to introduce a quantitative switch. Compared to the latter study, defining structural features and context specific mutations required for feature acquisition, is more robust and provides a better roadmap to aid in the surveillance and monitoring of H5 HA strains with pandemic potential. In a broader context, the unique integrated approach applied here lay important framework for studying human adaptation of other avian influenza HAs.

Introduction

The highly pathogenic H5N1 influenza A virus subtype poses a global health concern. This is evident from it having already led to several localized outbreaks in humans with a high case fatality ratio (~60%) since 2003[135,136]. The H5N1 subtype, however, has not yet adapted to the human host and established sustained human-to-human transmission via respiratory droplets (or aerosol transmission). One of the key factors governing adaptation of virus to human host is the glycan receptor-binding specificity of its hemagglutinin (HA). The HA from avian subtypes typically binds to α2→3 sialylated glycans (or avian receptors)[137]. A hallmark feature of human-adapted subtypes such as H1N1, H2N2, and H3N2 is the "quantitative switch" in their binding preference to α2→6 sialylated glycan receptors (or human receptors), which is defined by high relative binding affinity to human receptors over avian receptors. This quantitative switch in receptor specificity has been shown to correlate with the respiratory droplet transmissibility of the pandemic H1N1 and H2N2 viruses in ferrets [138] [41] [37] [139]. Therefore, a necessary determinant for human adaptation of avian-adapted H5N1 sub-type is the acquisition of mutations that quantitatively switch HA's binding preference to human receptors[137] [140]. Structure and receptor complexes of the HA from different subtypes (H1, H2, H3, and H5) have shed light on the nature of mutations preferred by avian and human viruses [141] [142] [143] [144]. These studies showed that the residues at 226 and 228 are critical determinants of the receptor-binding specificity of H2 and H3 HA, with human viruses favoring L226 and S228 and avian viruses favoring Q226 and G228 [144] [142]. On the other hand, for H1 viruses, 190 and 225 are critical determinants of the receptor-binding specificity, with human viruses favoring D190 and D225 and avian viruses favoring E190 and G225[141]. Identifying mutations that switch glycan receptor-binding preference of H5 HA has been the focus of several previous studies [15,145–150]. Some of these studies include analyses of glycan receptor binding of H5 HAs with natural variations in the receptor-binding site (RBS) [150]. Other studies have mutated H5 HA to include either the "hallmark" changes for human adaptation of H2/H3 HA (Q226L and G228S or LS) and/or H1 HA (E190D, G225D, or DD). More recently, two studies by Imai et al. (2012)[151] and Herfst et al. (2012)[152] demonstrated that specific sets of mutations in HA from A/Vietnam/1203/04 (Viet04) and A/Indonesia/5/05 (Ind05) viruses, respectively, confer respiratory droplet viral transmission in ferrets to the viruses possessing these mutant H5 HAs. From these studies, it is evident that differences in genetic background (using natural H5N1 isolate versus laboratory re-assorted strain) and selection pressure strategies give rise to distinct sets of amino acid changes in Viet04 and Ind05 HAs that are associated with aerosol transmission in ferrets. None of the wild-type (WT) natural variants or mutant H5 HAs from the aforementioned studies, however, have shown a quantitative switch in binding to human receptor in a fashion characteristic of "pandemic" strain HAs (such as 1918 H1N1, 1958 H2N2, and 2009 H1N1)(Shown in Figure 3.1).

Sequences of H5 HA isolated after 2006 have diverged considerably from the prototypic strains Viet04 and Ind05. This sequence divergence has critical implications for identifying amino acid changes in the RBS required to quantitatively switch the binding preference of H5 HA to human receptors. In this scenario, an important unanswered question is how current H5 HA would quantitatively switch to human

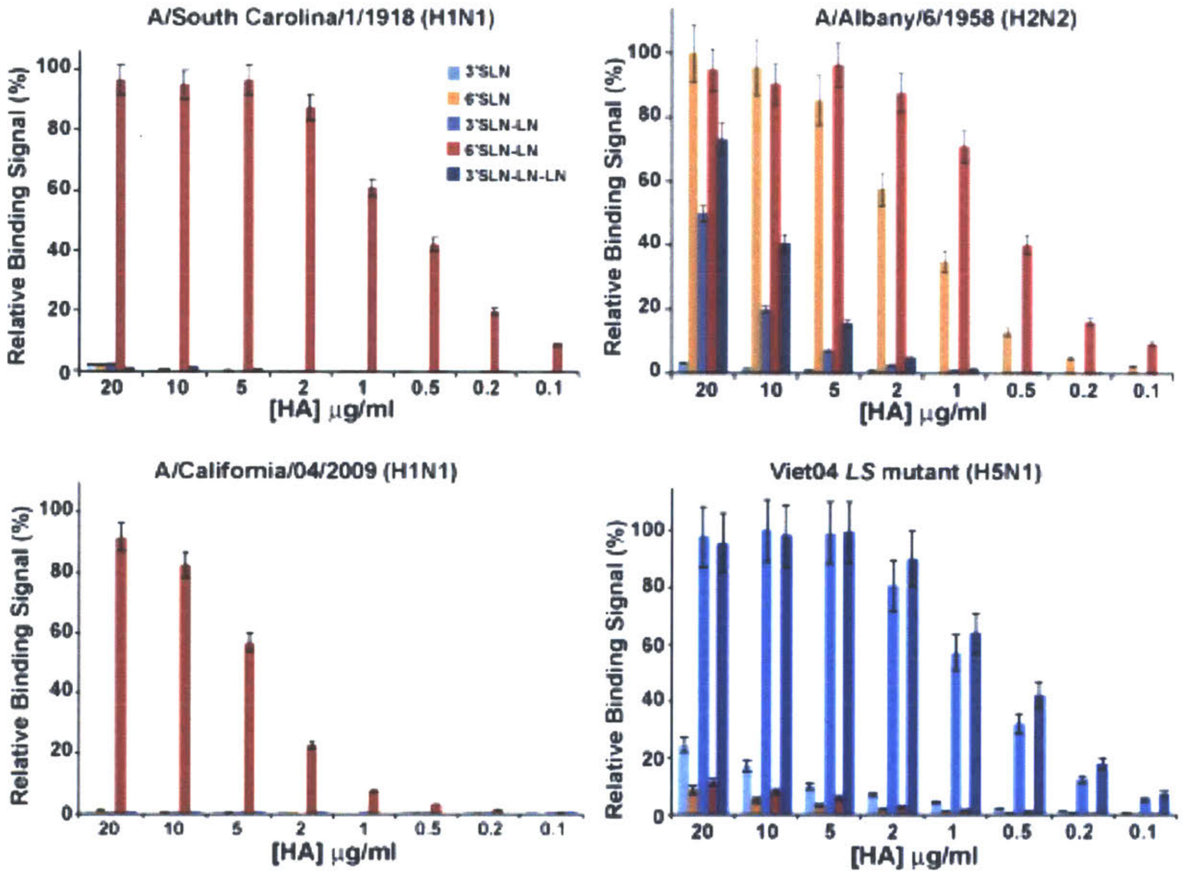


Figure 3.1 **Dose-Dependent Direct Glycan Array Binding of Prototypic Human-Adapted Pandemic HAs** HAs from prototypic human-adapted pandemic 1918 H1N1(top left), 1958 H2N2 (top right) and 2009 H1N1(bottom left) strains show specific high affinity binding to human receptors (6′ SLN-LN) with minimal to substantially lower affinity binding (relative to human receptor affinity) to avian receptors (3′ SLN-LN). On the other hand, introducing prototypic LS mutation on Viet04 does not quantitative switch its binding preference to human receptor (bottom right). Error bars were calculated based on normalized binding signals for glycan array assays done in triplicate for each HA sample.

receptor binding in the context of other molecular changes in its RBS due to sequence divergence from prototypic strains such as Viet04 and Ind05 [153] [149].

In this study, we have developed a distinct structural framework to systematically analyze the RBS of H5 HA from the perspective of structural topology of its glycan receptor, residues that interact with this receptor, and their inter-residue interactions in the RBS. Using this framework, we compared the RBS of H5 HA with that of H2 HA—its phylogenetically closest neighbor—to define molecular features that are critical for quantitative switching H5 HA binding to human receptors. Analysis of sequences for naturally evolving H5 HAs show that the different H5 clades have evolved to acquire distinct features that make them closer to human adaptation. We demonstrate that a subset of rapidly evolving and currently circulating H5 clades require as few as a single base pair change to quantitatively switch to human receptor binding. However, the acquisition of these distinct features in the various clades appears to be somewhat nuanced. We show here that amino acid changes that led to aerosol transmission of Viet04 and Ind05, when introduced in the HA of some of the currently circulating H5 clades, did not quantitatively switch these H5 HA binding to human receptors. Our study highlights the critical need to investigate RBS amino acid sequence divergence of H5 HA in the context of RBS molecular features to delineate H5N1 human adaptation.

Experimental Methods

Cloning, Baculovirus Synthesis, and Mammalian Expression and Purification of HA

Wild-type and mutant H5 HA sequences were codon optimized for insect cell expression and were synthesized at DNA2.0 (Menlo Park, CA). The synthesized genes were then subcloned into pAcGP67A plasmid, and baculoviruses were created using Baculogold system (BD Biosciences, San Jose, CA) according to the manufacturer's instructions. The recombinant baculoviruses were then used to infect suspension cultures of Sf9 cells cultured in BD Baculogold Max-XP SFM (BD Biosciences, San Jose, CA). The infection was monitored, and the conditioned media were harvested 3–4 days post-infection. The soluble HA from the harvested conditioned media was purified using Nickel affinity chromatography (HisTrap HP columns, GE Healthcare, Piscataway, NJ). Eluting fractions containing HA were pooled, concentrated, and buffer exchanged into 13 PBS pH 8.0 (GIBCO) using 100 kDa MWCO spin columns (Millipore, Billerica, MA). The purified protein was quantified using BCA method (Pierce).

The gene was codon optimized for mammalian expression, synthesized (DNA2.0, Menlo Park, CA), and sub-cloned into modified pcDNA3.3 vector for expression under CMV promoter. Recombinant expression of HA was carried out in HEK293-F FreeStyle suspension cells (Invitrogen, Carlsbad, CA) cultured in 293-F FreeStyle Expression Medium (Invitrogen, Carlsbad, CA) maintained at 37 degrees Celsius, 80% humidity, and 8% CO2. Cells were transfected with Poly-ethylene-imine Max (PEI-MAX, PolySciences, Warrington, PA) with the HA plasmid and were harvested 7 days postinfection. The supernatant was collected by centrifugation, filtered through a 0.45 mm filter system (Nalgene, Rochester, NY), and supplemented with 1:1,000 diluted protease inhibitor cocktail (Calbiochem filtration) and supplemented with 1:1,000 diluted protease inhibitor cocktail (EMD Millipore, Billerica, MA). HA was purified from the supernatant using His-trap columns (GE Healthcare) on an AKTA Purifier FPLC system. Eluting fractions containing HA were pooled, concentrated, and buffer exchanged into 13PBS pH7.4 using 100 kDa MWCO spin columns (Millipore, Billerica, MA). The purified protein was quantified using BCA method (Pierce, Rockford, IL). Both expression systems were used in this study. Importantly, no differences were observed in the glycan-binding properties of the HA derived from baculovirus when compared to that of the material derived from mammalian expression.

Homology Modeling of HA

A structural model of Alb58 HA trimer was built using the MODELER homology modeling software. To build the model, the solved crystal structure of A/Singapore/1/57 hemagglutinin with human receptor (PDB: 2WR7), which has 99% sequence identity in HA1 to Alb58, was used as a template. During modeling, the ligand (human receptor) was copied from the template structure into the model structure. The final model was minimized to release internal constraints.

Dose-Dependent Direct Binding of WT and Mutant HA

To investigate the multivalent HA-glycan interactions, a streptavidin plate array comprising representative biotinylated α2→3 and α2→6 sialylated glycans was used. 3'SLN, 3'SLN-LN, and 3'SLN-LN-LN are representative avian receptors. 6'SLN and 6'SLN-LN are representative human receptors. LN corresponds to lactosamine (Galβ1→4GlcNAc), and 3'SLN and 6'SLN, respectively, correspond to Neu5Acα2→3 and Neu5Acα2→6 linked to LN. The biotinylated glycans were obtained from the Consortium of Functional Glycomics through the resource request program. We have chosen a defined set of representative avian and human receptors given that the focus of our experimental studies is to quantitatively characterize relative binding affinity to human versus avian receptors and not to define specificity on the basis of number of human versus avian receptors using a larger array of glycans. The quantitative affinity defined using our defined glycan array platform has been used in several previous studies to correlate glycan-binding properties of HA with physiological properties of the virus such as respiratory droplet transmission in ferrets [154] [155] [156] and antibody response in mice[130]. Streptavidin-coated high-binding-capacity 384-well plates (Pierce) were loaded to the full capacity of each well by incubating the well with 50 ml of 2.4 mM of biotinylated glycans overnight at 4 degrees Celsius. Excess glycans were removed through extensive washing with PBS. The trimeric HA unit is composed of three HA monomers. The spatial arrangement of the glycans in the plate array favors binding to only one of the three HA monomers in the trimeric HA unit. Therefore, in order to specifically enhance the multivalency in the HA-glycan interactions, the recombinant HA proteins were pre-complexed with the primary and secondary antibodies in the molar ratio of 4:2:1 (HA:primary:secondary). The identical arrangement of four trimeric HA units in the precomplex for all the HAs permits comparison between their glycan binding affinities. A stock solution containing appropriate amounts of histidine-tagged HA protein, primary antibody (Mouse anti-6xHis tag IgG from Abcam), and secondary antibody (HRP conjugated goat anti- mouse IgG from Santacruz Biotechnology) in the ratio 4:2:1 was incubated on ice for 20 min. Appropriate amounts of pre-complexed stock HA were diluted to 250 ml with 1% BSA in PBS. 50 ml of this pre-complexed HA was added to each of the glycan-coated wells and incubated at room temperature (RT) for 3 hrs followed by the wash steps with PBS and PBST (13PBS + 0.05% Tween-20). The binding signal was determined based on HRP activity using Amplex Red Peroxidase Assay Kit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The experiments were done in triplicate. Minimal binding signals were observed in the negative controls, including binding of pre-complexed unit to wells without glycans and binding of the antibodies alone to the wells with glycans. The binding parameters, cooperativity (n), and Kd' for HA-glycan binding were calculated by fitting the average binding signal value (from the triplicate analysis) and the HA concentration to the linearized form of the Hill equation:

$$\log\left(\frac{y}{1-y}\right) = n * \log\left([HA] - \log\left(Kd'\right)\right.$$

where y is the fractional saturation (average binding signal/maximum observed binding signal). In order to compare Kd' values, the values reported in this study correspond to the appropriate representative avian (3'SLN-LN or 3' SLN-LN-LN) and human (6'SLN-LN) receptors that gave the best fit to the above equation and the same slope value (n value is -1.3). As noted above, there were no differences in the glycan-binding properties for HA derived from baculovirus when compared to that of HA produced via mammalian expression (unpublished observation).

Binding of WT and Mutant HAs to Human Tissue Sections

Paraffinized human tracheal and alveolar (US BioChain) tissue sections were deparaffinized, rehydrated, and incubated with 1% BSA in PBS for 30 min to prevent nonspecific binding. HA was pre-complexed with primary antibody (mouse anti-6xHis tag, Abcam) and secondary antibody (Alexa Fluor 488 goat anti-mouse, Invitrogen) in a molar ratio of 4:2:1, respectively, for 20 min on ice. The tissue binding was performed over two different HA concentrations (40 mg/ml and 20 mg/ml) by diluting the pre-complexed stock HA in 1% BSA- PBS. Tissue sections were then incubated with the

HA-antibody complexes for 3 hr at RT. The tissue sections were counterstained by propidium iodide (Invitrogen; 1,100 in TBST). The tissue sections were mounted and then viewed under a confocal microscope (Zeiss LSM 700 laser scanning confocal microscopy). Sialic-acid-specific binding of HAs to tissue sections was confirmed by loss of staining after pretreatment with Sialidase A (recombinant from *Arthrobacter ureafaciens*, Prozyme). This enzyme has been demonstrated to cleave the terminal Neu5Ac from both Neu5Aca2→3Gal and Neu5Aca2→6Gal motifs. In the case of sialidase pretreatment, tissue sections were incubated with 0.2 units of Sialidase A for 3 hr at 37 degrees Celsius prior to incubation with the proteins. Pretreatment of human tracheal and alveolar tissue sections with Sialidase A resulted in complete loss of HA staining.

Capturing Network of RBS Residues

The coordinates of the H5 HA-avian receptor and Alb58 HA-human receptor complex were uploaded into the PDBePISA server (http://www.ebi.ac.uk/ msd-srv/prot_int/pistart.html) to determine key residues in the HA RBS that make contact with the corresponding glycan receptor (interface cutoff of 30% was used). For these residues, their environment was defined using a distance threshold of 7 Angstroms , and the contacts, including putative hydrogen bonds (which include water-bridged ones), disulfide bonds, pi bonds, polar interactions, salt bridges, and Van der Waals interactions (nonhydrogen), occurring between pairs of residues within this threshold distance were computed as described previously [157]. These data were assembled into an array of eight atomic interaction matrices. A weighted sum of the eight atomic interaction matrices was then computed to produce a single matrix that accounts for the strength of atomic interaction between residue pairs within the RBS, using weights derived from relative atomic interaction energies [157]. The inter-residue interaction network calculated in this fashion generates a matrix that describes all the contacts made by critical RBS residues with spatial proximal neighboring residues in their environment. For each element i, j is the sum of the path scores of all paths between residues i and j. The degree of networking score for each residue was computed by summing across the rows of the matrix, which was meant to correspond to the extent of "networking" for each residue. The interactional relationship between critical RBS residues and their environment is represented using a two-dimensional open connectivity network diagram (RBSN diagram). The degree of networking score was normalized (RBSN score) with the maximum score for each protein so that the scores varied from 0 (absence of any network) to 1 (most networked). Although previous studies [152] [151] have reported mutations in stalk and other regions of HA outside the RBS, the RBSN does not include these residue positions, and hence, any amino acid changes in these positions would not affect the glycan receptor- binding property of HA.

Sequence Analysis of H5 HA and Estimation of Key Features

A total of 6,014 H5 HA sequences were downloaded from the EpiFlu database. From this, only those sequences that had complete coding regions, including start and stop codons, were considered. In order to avoid estimation errors due to multiply represented sequences, all groups of identical sequences in the data set were represented by the oldest sequence in the group. The remaining 2,959 sequences were ordered by isolation time and aligned, and the occurrence rate of each feature (defined as the percent fraction of sequences from a given year that contains that feature) was calculated.

Phylogeny Tree Construction and Calculation of Sequence-Based Distance Measure for RBS Analysis

Phylogeny tree was constructed for 2,959 H5 HA amino acid sequences by neighbor-joining method using MEGA 5.10 software (http://www. megasoftware.net/). The branches were colored based on the different features present (Table 3.1) using the Rainbow Tree software (http://www.hiv.lanl.gov/content/sequence/RAINBOWTREE/rainbowtree.html).

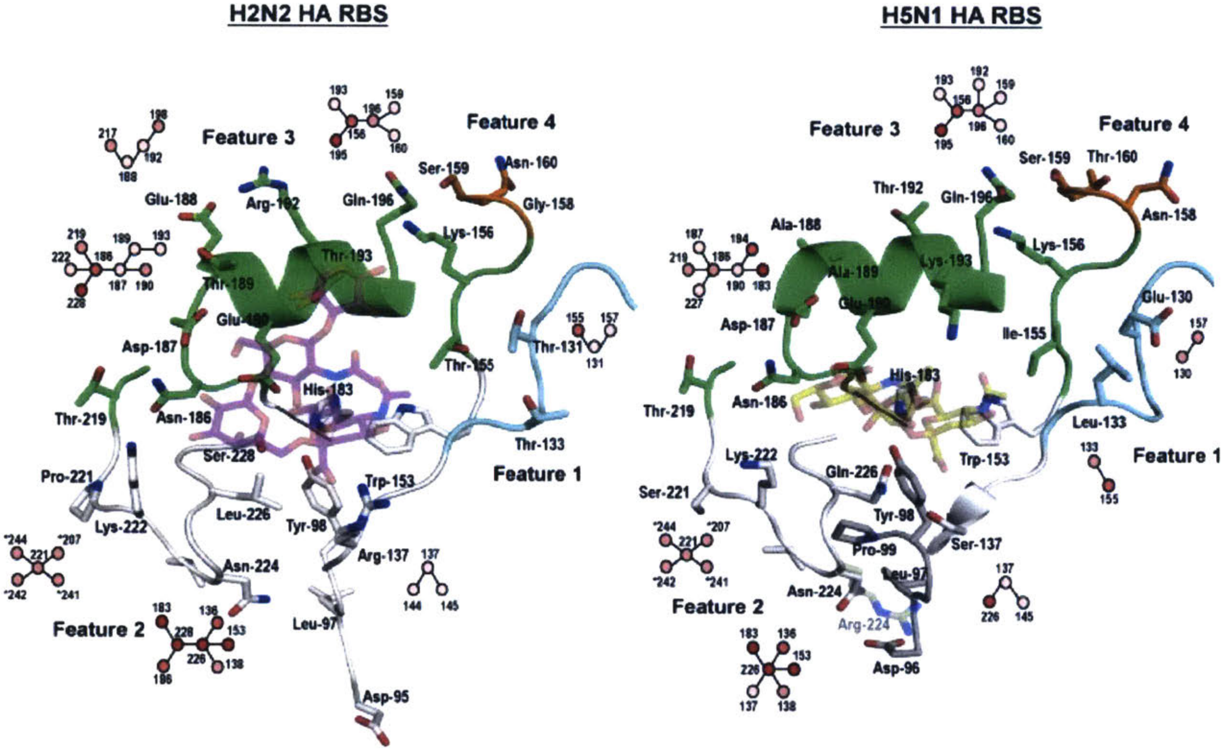


Figure 3.2 **Comparing Molecular Features in the RBS of H5 and H2 HA**. The RBS of H5 and its phylogenetically closest H2 HA is rendered as a cartoon. The key residues are labeled, and their side chains are shown in stick representation. The human receptor and avian receptor are shown with carbon atoms colored in magenta and yellow, respectively, in the stick representation at 50% transparency. The four key features are indicated based on coloring the carbon atom in different colors. The 130 loop (feature 1) is shown in cyan. The base of the RBS (feature 2), including residue positions 136–138, 153, and 221–228, is shown in gray. The side chain of Arg in the 224 position is shown at 50% transparency. The top of the RBS (feature 3), including the 190 helix and residue positions 155, 156, and 219, is shown in green. The glycosylation sequon at the 158 position (feature 4) in H5 HA is shown in orange. The RBSN diagram is shown adjacent to the residue positions in that network. The circular nodes are colored according to their RBSN score (pink representing a low score to bright red representing a high score) and their connectivity to other nodes. The asterisk next to residue positions indicates that these positions belong to the adjacent HA1 domain in the HA trimer. The HA structure images were generated using Pymol (http://www.pymol.org/).

Results

Defining Molecular Features for H5 HA RBS

We began by systematically analyzing the RBS of H5 HA from the perspective of structural topology of its natural avian receptor and that of the amino acids in the RBS. Previously, we had developed a framework to distinguish binding of HA to avian and human receptors on the basis of the three-dimensional structural topology of these receptors[15]. When bound to HA, the avian receptor sampled a conformational space that resembles a cone (thus the term cone-like topology was used to describe this receptor). The majority of contacts of H5 HA (using Viet04 crystal structure[146] [147]) with avian receptor adopting a cone-like topology involve the Neu5Acα2→3Gal motif. The key amino acids in the H5 HA RBS involved in this interaction predominantly lie in the base of the RBS, involving residues Ser-136 in 130 loop, Trp-153, Ile-155 in the 150 loop, Lys-222, and Gln-226 in the 220 loop with specific additional contacts from Glu-190, Lys-193, and Leu-194 in the 190 helix at the top of the RBS (Figure 3.2).

There are several human-adapted HAs, including various seasonal and pandemic strains from H1N1, H3N2, and the pandemic H2N2 subtypes. Based on the phylogenetic "closeness" of H5 HA to H2 HA (Figure 3.3), we selected human-adapted H2N2 HA (A/Albany/6/58 or Alb58) bound to human receptor to contrast with the structural model of H5 HA bound to avian receptor[146] [147]. We chose Alb58 as a representative H2N2 strain because it is a prototypic pandemic strain, and we have already extensively characterized its quantitative glycan receptor binding and phenotypic properties (such as aerosol transmissibility) [138] [139]. As the X-ray crystal structure of Alb58 HA is not available, we constructed a homology-based model for this HA using the template crystal structure of another human-adapted H2 HA (A/Singapore/1/57) (co-crystallized with human receptor), which has high sequence identity to Alb58 HA [142].

The human receptor bound to HA samples a larger conformational space that resembles a fully closed to fully open umbrella (and thus we have used the term umbrella-like topology to define this receptor). There are two regions in the umbrella-like topology of the human receptor: the base region composed of Neu5Acα2→6Galβ1→ motif and an extension region comprising sugar residues beyond this motif (typically GlcNAcβ1→3Galβ1→). These two regions span a wider range of interacting amino acids in the H2 HA RBS. A comparison of H5 HA bound to avian receptor in cone-like topology and H2 HA bound to human receptor in umbrella-like topology showed four key differences (Figure 3.2). First, the composition of the 130 loop of H2 HA is different from H5 HA in that there is a deletion in this loop. The deletion in the 130 loop in H5 HA relative to H2 HA critically influences the 130 loop. Second, amino acids in the "base" of the RBS (such as those in the 130 loop at positions 136–138 and the 220 loop at positions 219–228) are different. Third, the "top" of the RBS primarily comprising the "190 helix" (residues 188–196) that interacts with the extension region of human receptor in H2 HA is different (specifically at positions 188, 189, 192, and 193). Fourth, position 158 is glycosylated in H5 HA, but not in H2 HA. Glycosylation at this site could potentially interfere with the extension region of human receptor [147].

To determine what mutations are needed to overcome the above differences and so that H5 HA can switch its specificity to human receptor, we developed a metric (RBS network or RBSN) to capture the network of interactions between the critical residues in the RBS and other residues in their close spatial environment that make contact with the glycan (see methods). The higher the network score of an amino acid within the RBS, the more structurally constrained it is to be mutated. For example, residues that make critical contacts with sialic acid such as Phe-98, Trp-153, and His-183 are highly networked (RBSN scores > 0.7) and, hence, are less constrained to mutate (Figure 3.4).

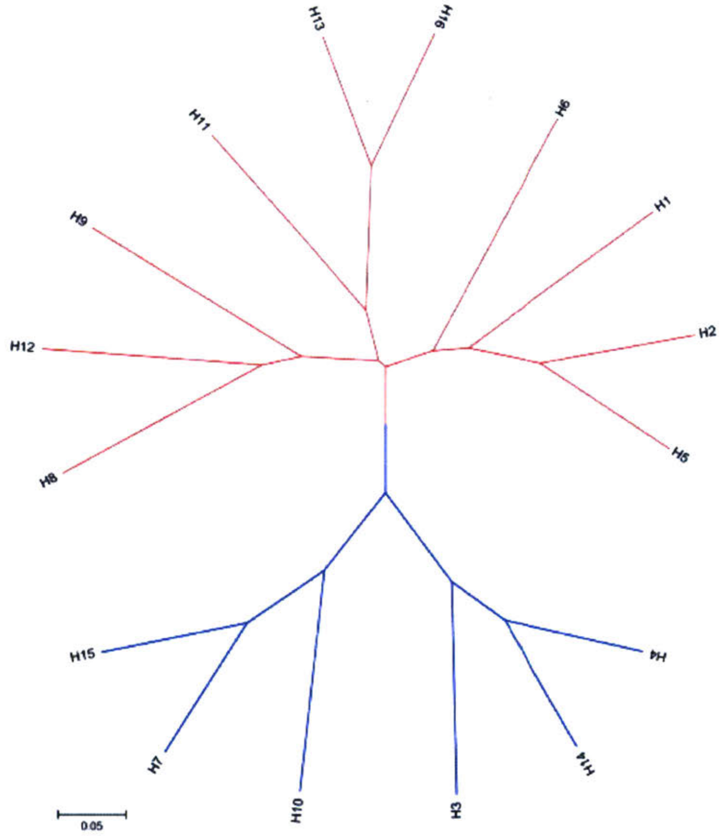


Figure 3.3 **Phylogenetic Tree of Different HA Subtypes.** Branches leading to group 1 & 2 HAs are labeled and colored in red and blue, respectively. Closely related subtypes are located on branches close to one another.

On the other hand, residues that are at the interface of the RBS and antigenic sites such as in the 130 loop, 190 helix, and 220 loop are poorly to moderately networked (RBSN scores < 0.15) and can readily undergo mutations as a part of antigenic drift. Thus, we define the term—molecular feature (for each of the four differences)—that incorporates glycan topology, HA residues involved in the binding, and their inter-residue interaction network in the RBS. We demonstrate that such an approach provides a robust framework to investigate amino acid mutations (and hence matching features) that quantitatively switch binding of H5 HA to human receptors in a manner similar to what has been observed for pandemic HAs. Amino acid changes related to feature 1 involve altering the length of the 130 loop specifically by introducing a deletion, which, in turn, affects the RBSN diagram involving 131, 133, and 155 positions. Residues at positions 131 and 133 had low RBSN scores (<0.04) in H5 HA and therefore could be readily mutated in context of the 130 loop deletion such that it contributes to human receptor specificity. Changing the base of the RBS (feature 2), which plays a critical role in governing glycan receptor specificity, involves alteration of a combination of residue positions in the 130 loop and 220 loop. In H5 HA, Gln-226 plays a critical role in contacts with Neu5Acα2→3Gal→motif of avian receptor, and Ser-137 and Gln-226 are involved in the inter-residue interaction network. Conversely, in H2 HA, the corresponding Leu-226 and Arg-137 are not related. Arg-137 and Ser-228 in H2 HA provide additional stabilizing contacts with sialic acid. Therefore, one way to match feature 2 (for human receptor binding) involves changing residues at 137 and 226 positions in H5 HA. Residue position at 137 is readily mutable given its low RBSN score in H2 and H5 HA (~0.01).

However, residue at 226 has a much higher RBSN score in H2 and H5 HA (>0.25). Making changes to this residue therefore would also involve making other changes—specifically, changing Gly-228/Ser in addition to Ser-137/Arg mutation. Although Gln-226/Leu mutation governs switch in contacts from Neu5Acα2→3Gal→ to Neu5Acα2→6Gal→motif, Ser-137/Arg and Gly-228/Ser mutation provides additional stabilization to the 130 and 220 loop at the base of the H5 RBS from the standpoint of inter-amino acid networking and improved contacts with glycan receptor [146] [147]. This stabilization can also be accomplished by mutation Asn-224 to Lys or Arg (naturally observed in some pandemics), as this would enhance its inter-amino acid interaction network with Asp-96, Leu-97, and Pro-99 (Figure 3.2). Therefore, feature 2 can also be matched by fewer mutations at 224 and 226 positions. The RBSN diagram of the residue at the 221 position in H5 HA is identical to that in H2 HA, although this position has a Ser in H5 HA and a Pro in H2 HA (Figure 3.2). It is likely for the Pro to govern the conformation and side-chain orientation of the adjacent Lys-222 residue, which plays a key role in making contacts with the human receptor. Therefore, changing Ser-221/Pro in H5 HA would permit more optimal conformation of the 220 loop for contact with the human receptor. On the other hand, mutations at positions 188, 189, 192, and 193 and the RBSN diagrams depicting their interaction networks (Figure 3.2) will impact feature 3.

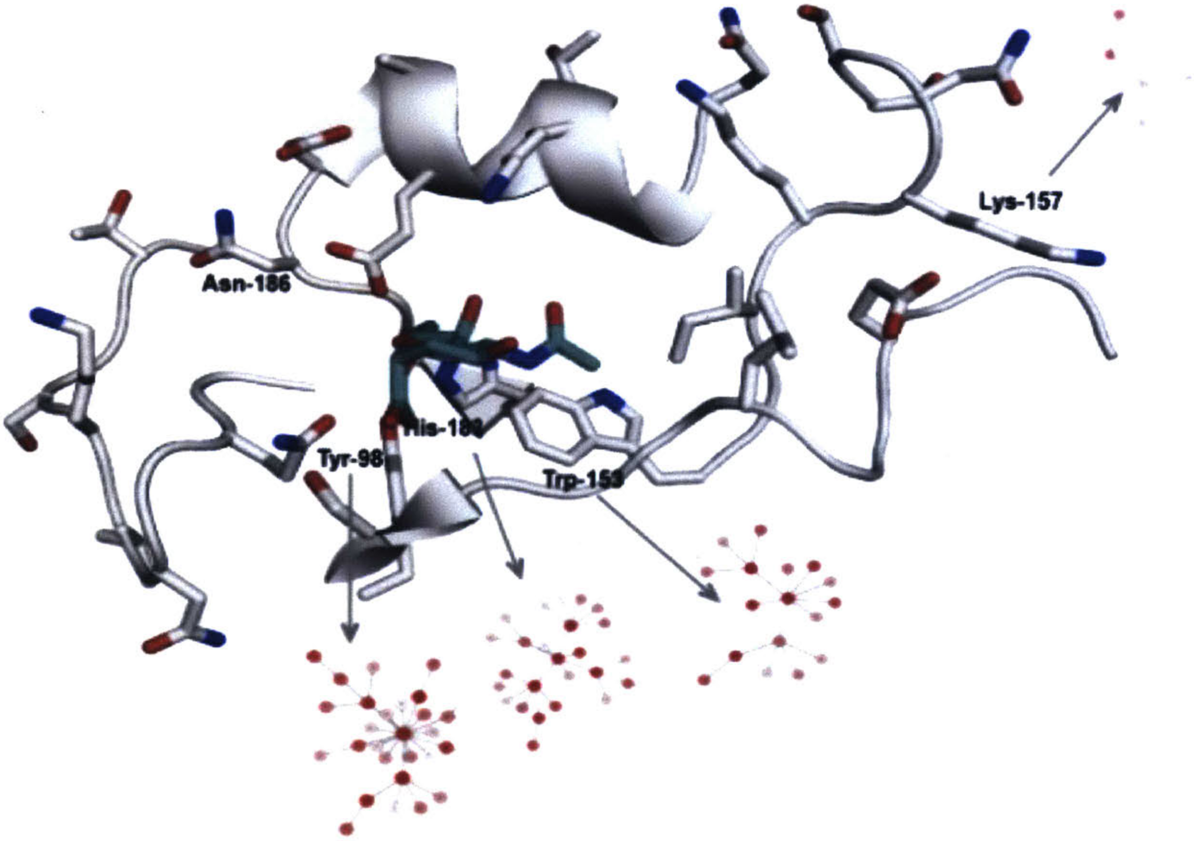


Figure 3.4 **RBSN Diagram of Representative Residues**. The sialic acid anchoring residues such as Tyr-98, Trp-153 and His-183 are highly networked and hence structurally constrained to mutate whereas a residue in antigenic site such as Lys-157 is moderately networked and hence can readily undergo mutations as a part of antigenic drift.

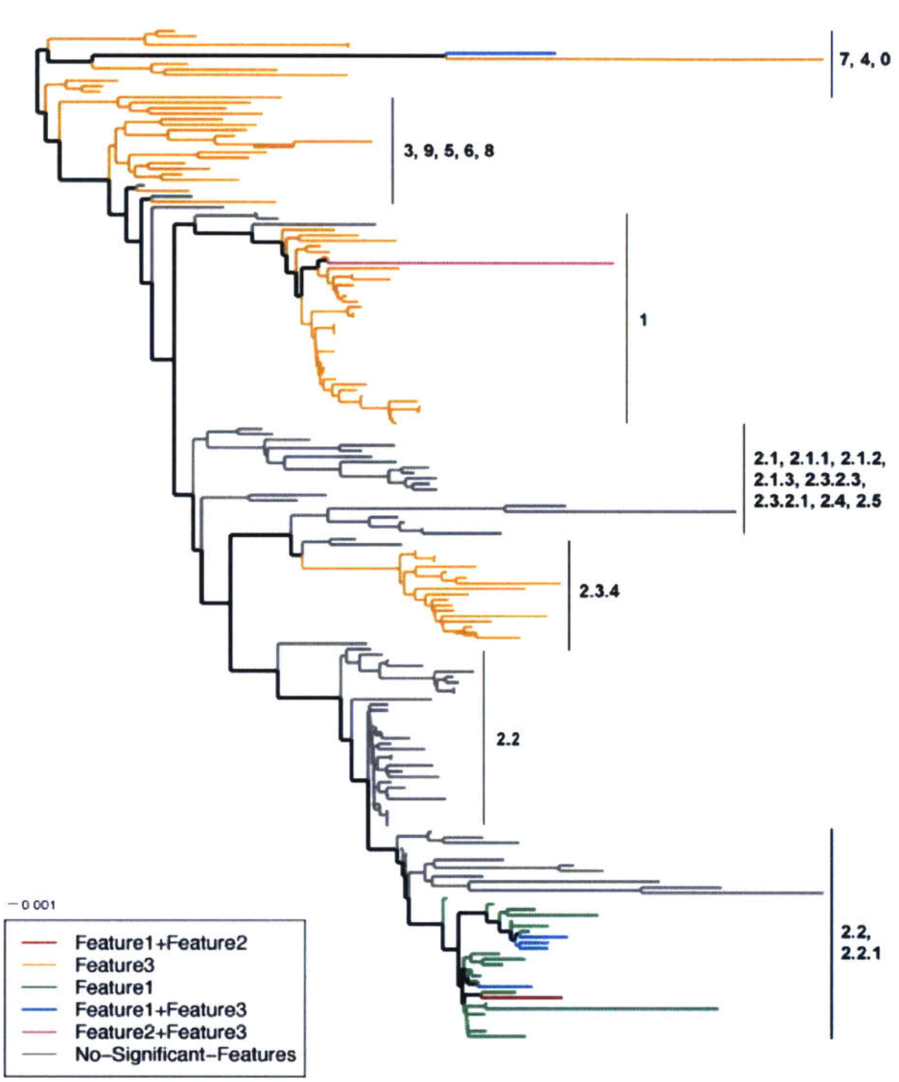


Figure 3.5 **Phylogenetic Tree of Representative A/H5N1 HA1 Protein Sequences Showing Clustering of Sequences Based on Antigenic Clades**. The sequences are color coded by the features observed in the HA1 domain. Scale bar represents 0.001 amino acid substitutions per site. Feature 1 has evolved in clade-2.2, -2.2.1, and -7 strains after 2006, whereas feature 2 has evolved in clade-1 and -2.2.1 after 2007. Feature 3 has evolved in clade-2.2 and -7 strains. Some of the currently circulating clades (1, 2.1.3, 2.2, 2.2.1, 2.3.2, 2.3.4, and 7) have acquired mutations to match multiple RBS features and hence appear to be closer to human adaptation. Among these H5 HAs, the relative order of dominant circulating clades found to have acquired multiple features includes clade-2.2.1, -2.2, -1, and -7. A subset of clade 2.2.1 and clade 7 has acquired amino acid changes in the RBS base toward matching feature 1 and/or feature 2.

The residues Ala-188, Ala-189, and Thr-192 in H5 HA do not have any inter-residue contacts with other residues in the RBS. However, residues Glu-188, Thr-189, and Arg-192 are involved in multiple interaction networks. Therefore, amino acid changes in positions 188, 189, 192, and 193 (hence altering feature 3) of H5 HA RBS would also impinge on its human receptor-binding property. Given that the RBSN scores of all these residue positions are low (<0.1) in H5 HA, they are readily mutable. Finally, removal of glycosylation sequon at position 158 relates to feature 4. This could be accomplished by mutating Asn-158 to a residue such as Asp or by mutating Thr-160 to Ala.

Among these four distinct features, feature 2 critically influences the base of the RBS, which plays an important role in distinguishing contacts between H5 and Neu5Acα2→3Gal→ motif (avian receptor) and H2 HA and Neu5Acα2→6Gal→ motif (human receptor). The difference in the 130 loop length (feature 1) would potentially affect the base of the RBS, which, in turn, is critical for glycan-receptor specificity. Changes related to feature 2 to achieve the necessary glycan receptor specificity therefore would critically depend on the 130 loop length (feature 1). Conversely, feature 3, which involves the 190 helix, plays an important role in contacts with avian receptor in H5 (specifically Lys or Arg at 193 position) and in contacts with extension region of human receptor. Therefore, amino acid changes in H5 HA related to feature 3 would reduce binding to avian receptor and enhance contact with extension region of human receptor. Feature 4 relates to the state of glycosylation at position 158, which impinges on the interaction with the extension region of human receptor [147].

Analyzing Natural Sequence Evolution of H5 HA from the Perspective of RBS Features

Having described and contrasted the H5 HA RBS features (henceforth referred to as RBS features) in the context of H2 HA RBS, we then investigated the acquisition of these features in the context of sequence evolution of H5 HA over time instead of searching for specific hallmark human-adaptive mutations described in previous studies [158] [159] [160] [147]. A phylogeny map of A/H5N1 HA1 sequences was constructed, and the branches were color-coded based on the features present (Figure 3.5). Feature 4 has been observed in almost all of the clades; however, amino acid changes characteristic of features 1, 2, and 3 were observed more recently. Many of the clades, including currently circulating clade 1, clade 2.2, clade 2.2.1, and clade 7, have acquired amino acid changes characteristic of one or two of features 1, 3, and 4. However, only a subset of the rapidly evolving and currently circulating clades 2.2.1 and 7 have acquired amino acid changes— which are critical features of the RBS base—to match feature 1 and/or part of feature 2. In the context of the key structural features of HA RBS, the deletion in the 130 loop with a concurrent loss of glycosylation (features 1 and 4) in the same HA was the most critical change observed in the evolution of 2.2.1 H5 HA since 2007. A recent study [160] by Russell compared the sequence evolution of H5N1 simply in the context of hallmark changes reported by the two H5N1 transmission studies involving Viet04 [151] and Ind05 [152] indicated that clade 2.2.1 viruses were most similar to human-transmissible virus. In this earlier analysis, clade 7 does not appear to be significant. On the other hand, apart from a subset of clade 2.2.1 strains that have acquired one or more RBS features, many clade 2.2.1 strains have not yet acquired any of the RBS features, hence they are almost indistinguishable from non-circulating clades (3, 5, 6, 8, and 9). Taken together, the above findings suggest that acquisition of these distinct features in the various clades appears to be somewhat nuanced and that subsets of the currently circulating H5 HA strains have not only diverged considerably from older human isolates (such as Viet04) but have also acquired the key molecular RBS features necessary for human receptor preference.

Design and Validation of Amino Acid Changes Required to Quantitatively Switch Receptor Specificity of Currently Circulating H5 HA

Based on our analyses of the rapidly evolving and currently circulating H5 HAs, we chose representative examples from a subset of clades 2.2.1 and 7, which have already acquired amino acid changes to match one or two of the RBS features, to validate the framework and predictions. We chose A/chicken/Vietnam/NCVD-093/08 (ckViet08; clade 7 HA that acquired changes in feature 3), A/Egypt/N03450/2009 (Egy09; clade 2.2.1 HA that acquired changes in features 1 and 4), and A/duck/Egypt/10185SS/2010 (dkEgy10; another clade 2.2.1 HA that acquired changes in features 1, 4, and part of feature 2 based on Asn224/Lys natural mutation). On these HAs from these distinct recent H5 isolates, we introduced amino acid changes to match one or more of the remaining features. We expressed recombinant WT and mutant forms of these HAs and evaluated their glycan-binding properties in dose-dependent direct glycan array assay. A summary of WT and mutant HAs with their corresponding RBS features and glycan-binding properties is shown in Table 3.1. The details of the amino acid changes to match RBS features are summarized in Table 3.1. The dose-dependent glycan array data for the WT and mutant H5 HAs are shown in Figure 3.6. From Figure 3.6 and Table 3.1, it is evident that avian receptor binding of WT H5 HAs that have already acquired one or more of the RBS features is not as high as (specifically binding to 3'sialyl-lactosamine (3'SLN)) what we have observed for other avian-adapted HAs [15] [139]. In the case of ckViet08 (Figure 3.6A), the natural acquisition of changes in the 190 helix (feature 3), especially Met-193, is consistent with lowering of avian receptor binding because Lys or Arg at 193 in other H5 HAs (which have not acquired feature 3) make optimal contact with the avian receptors. In the cases of Egy09 and dkEgy10 (Figure 3.6C and E), the130 loop deletion (feature 1) does appear to also have some detrimental effects on avian receptor binding.

	Features of Human Adaptation					
Human Receptor→	**Base (Neu5Acα2→6Galβ1→)**		**Extension (→4GlcNAcβ1→3Galβ1→4→)**		**Glycan-Binding Specificity**	
	F1 (Δ130 loop)	**F2 (RBS base)**	**F3 (190 helix)**	**F4 (no 158 glyco)**	**α2→3**	**α2→6**
Egy06 (E3.0)				**D158/A160**	**++++**	**n.b.**
E3.1		**Q226L/G228S**		D158/A160	+++	+++
ckViet 08 (V4.0)			**K192/M193**		**+++**	**+**
V4.2		**Q226L/G228S**	K192/M193	**T160A**	++	++
V4.3		**N224K/Q226L**	K192/M193	**T160A**	n.b.	n.b.
V4.4	**L133Δ**	**N224K/Q226L**	K192/M193	**N158D**	+	+++
V4.5	**E130Δ/L133T**	**S137R/Q226L/ G228S**	**N187D**/K192/**M193T**	**T160A**	+	++++
Egy09 (E4.0)	**Δ133**			**A160**	**+++**	**n.b.**
E4.1	Δ133	**S137R/Q226L/ G228S**		A160	+	++
E4.2	**A131T**/Δ133	**S137R/S221P/ Q226L/G228S**	**R193T**	A160	+	+++
E4.3	Δ133	**N224K/Q226L**		A160	n.b.	+++
dkEgy 10 (E5.0)	**Δ133**	**K224**		**D158/A160**	**++**	**n.b.**
E5.1	Δ133	K224/**Q226L**		D158/A160	n.b.	+++

Table 3.1 **The key amino acids in WT H5 that contribute to acquisition of the corresponding feature**. These are indicated by WT amino acid (one letter code) followed by the HA position, whereas mutations (shown in **bold red**) introduced to match the features are indicated by WT amino acid followed by the HA position and the mutant amino acid. The blank feature columns indicate absence of H2-like RBS feature in the H5 HAs. The apparent avian and human receptor-binding affinities are indicated in the α2→3 and α2→6 columns in which highest = ++++, high = +++, moderate = ++, and low = +, and no observable binding = n.b. column.

We next sought to understand the relationship between molecular features in recent H5 HAs and the hallmark amino acid changes in Viet04 and Ind05 HA that conferred respiratory droplet transmission to viruses having these mutant HAs [152] [151]. The LS mutation and loss of glycosylation sequon at the 158 position were the RBS mutations previously reported for Ind05 [152]. The LS amino acid mutations only partially matched RBS base feature 2. Introducing the LS mutation with loss of 158 glycosylation sequon on ckViet08 (V4.2) showed some increased human receptor binding while retaining most of the avian receptor binding and, therefore, did not quantitatively switch this mutant HA binding to human receptors. Introducing the 130 loop deletion and amino acid changes in addition to LS and loss of glycosylation at 158 (see V4.5 in Table 3.1) completely switched binding to the human receptor. Therefore, the LS mutation alone is not sufficient to completely match (feature 2). The Asn-224/Lys/Gln226/Leu and loss of the 158-glycosylation sequon were the RBS mutations previously reported for Viet04 [151]. As an example, introducing these mutations on ckViet08 (V4.3) wholly abolished binding to both avian and human receptors. However, introducing the deletion in the 130 loop of V4.3 resulted in a mutant (V4.4) that quantitatively switched binding to human receptors (Figure 3.6B) (apparent binding affinity constant [Kd'] for 6'SLN-LN ~100 pM), highlighting the critical importance of the RBS base (feature 2) and the 130 loop deletion (feature 1) in modulating human receptor-binding specificity. Jointly, these results clearly demonstrate that the same amino acid mutations that led to aerosol transmission of Viet04 and Ind05 are not sufficient to quantitatively switch current circulating H5 HA binding to human receptors. Egy09 already acquired amino acid changes to match features 1 and 4. Similar to what was observed with ckViet08 V4.2 mutant, introducing the LS mutations in the RBS base (partially matches feature 2) even with the 130 loop deletion in Egy09 (E4.1) did not quantitatively switch its binding to human receptor. Instead, matching feature 2 by introducing the Asn-224/Lys/Gln- 226/Leu mutations on Egy09 (E4.3) quantitatively switched its binding to the human receptor (Kd' ~50 pM) (Figure 3.6C and D). The dkEgy10 is one of a kind because it has evolved to be the closest matching RBS features 1, 2, and 4. The glycan receptor binding of dkEgy10 HA (Figure 3.6E) shows that the WT HA still predominantly binds to avian receptors (3'SLN-LN and 3'SLN-LN-LN with minimal binding to 3'SLN and human receptors). The predominant avian receptor-binding property of dkEgy10 can be attributed to the 226 position, which still has a Gln (and not Leu) that is needed to match feature 2. In fact, introducing this single Gln226/Leu amino acid mutation (which involves a single base pair mutation) on dkEgy10 (E5.1) quantitatively switched its binding to human receptors (Kd' ~100 pM) (Figure 3.6F). The above mutant H5 HAs that show a switch in receptor preference demonstrate the necessary apparent binding affinity to human receptor

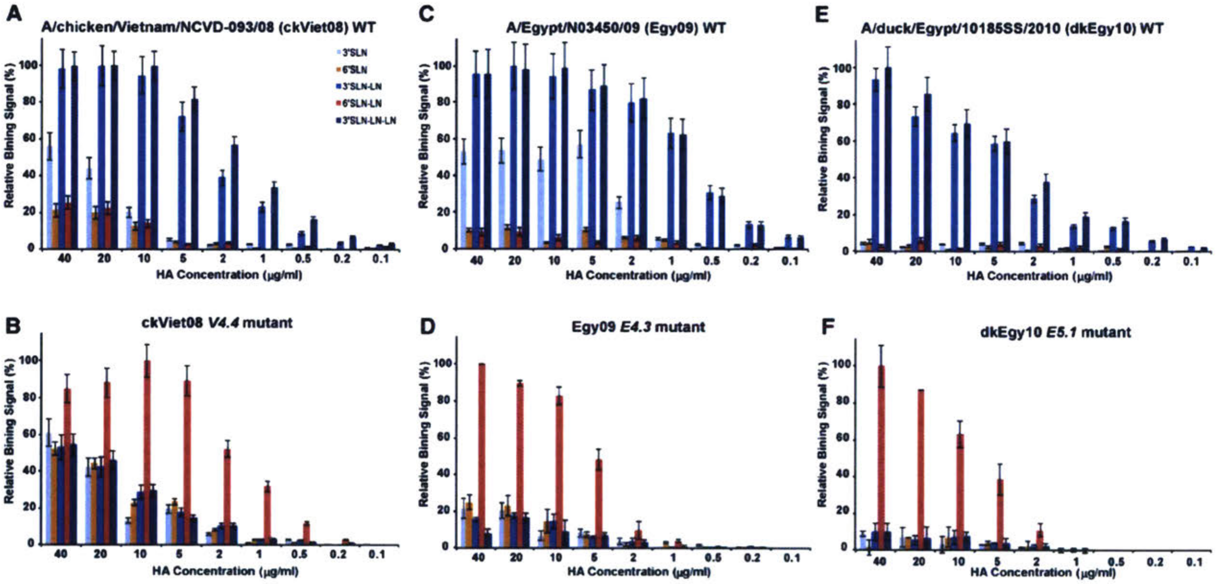


Figure 3.6 **Glycan Receptor Binding of WT and Mutant of H5 HAs** HA. (A–F) Dose-dependent direct binding of WT ckViet08, Egy09, and dkEgy10 is shown in (A), (C), and (E), respectively, and that of the V4.4, E4.3, and E5.1 mutants are shown in (B), (D), and (F), respectively. See also Table 3.1 for descriptions of the mutants. Error bars were calculated based on normalized binding signals for glycan array assays done in triplicate for each HA sample.

(similar to the 2009 H1N1 pandemic HA) and hence pass the threshold for potential aerosol viral transmission.

To extend the quantitative characterization of glycan receptor specificity to binding to physiological glycan receptors in human respiratory tissues, we analyzed the binding of dkEgy10 WT and E5.1 mutant on human tracheal and alveolar tissues. Studies previously have demonstrated that the apical surface of human tracheal tissue predominantly expresses human receptors, and the human alveolar tissue predominantly expresses avian receptors [15] [161] [162] [163]. Consistent with the quantitative switch to human receptor binding, E5.1 mutant showed the expected staining of the apical surface of the human tracheal tissue section, whereas the WT dkEgy10 showed extensive staining of the human alveolar section (Figure 3.7A). The above results together underscore the importance of delineating key structural RBS features in H5 in conferring the quantitative switch in binding to human receptors (Figure 3.6B). The base of the RBS captured by feature 2, which plays a critical role by making contacts with the terminal sialic acid linkage, appears to be a key determinant. Our data show (and are consistent with RBSN) that a combination of two amino acid changes, Asn-224/Lys and Gln-226/Leu in the RBS base, is able to match feature 2 effectively in specific H5 clades when compared to the LS combination in which additional mutations match the same feature. Our results also support a critical role for the "130 loop length" in augmenting the RBS base for optimal contacts with the human receptor because changes to the RBS base alone with Asn-224/Lys/Gln-226/Leu completely abolished binding of ckViet08 V4.3 mutant. Many strains in the H5 clade 2.2.1 do not possess the "130 loop" deletion (feature 1). In these strains, matching the critical feature 2, with feature 4 already matched, does not confer a switch in H5 receptor specificity. A representative clade 2.2.1 example (A/Egypt/2786-NAMRu3/06 [Egy06] and its mutant [E3.1]) is shown in Table 3.1. These strains are further away from human adaptation, and it cannot be generalized, therefore, that all clade 2.2.1 strains are closest to human adaptation. The 130 loop deletion has been naturally acquired by a subset of clade 2.2.1 HA and shows a lowering in binding to avian receptors as observed in both Egy09 and dkEgy10 strains. Our structural analysis and data also indicate that involvement of features 3 and

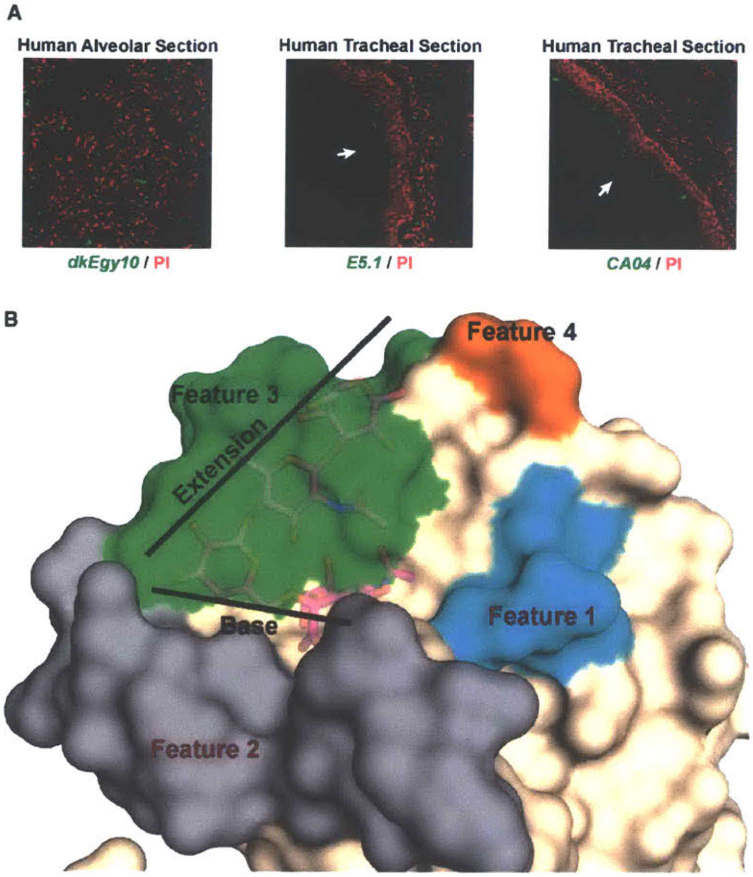


Figure 3.7 **Physiological Glycan Receptor Binding and Summary of RBS Features in Current H5 HAs**. In (A), the left panel shows staining of human alveolar section by dkEgy10. The middle panel shows staining of human tracheal tissue section by the E5.1 mutant. The right panel shows staining of human tracheal tissue section by CA04. For all the tissue sections, the HA staining is shown in green against propidium iodide shown in red. Apical surface of trachea is indicated by a white arrow (B). Surface rendering of H5 RBS with the region corresponding to features 1–4 is colored cyan, gray, green, and orange, in that order. The human receptor is shown as a stick. Features 1 and 2 are colored dark red to indicate their necessary requirement for human adaptation of H5 HA in the context of its current sequence evolution

4—which impact contacts with the extension region of the human receptor—in human receptor switch depends on clade-specific H5 HA RBS. Consequently, in the context of the current HA sequence evolution of rapidly evolving H5N1, features 1 and 2 appear to be necessary and additionally appear to be sufficient to match either feature 3 or 4 for human adaptation of distinct H5 HA clades.

Conclusion

In summary, we have developed a distinct approach to define the molecular features that characterize RBS of HA and have used these features to compare RBS of H5 HA with that of its closest phylogenetic neighbor—pandemic H2 HA. This network approach permits us to understand how amino acid changes (resulting from the extensive sequence divergence of H5 HA as a part of its natural evolution) in the RBS modulate glycan-binding specificity. Using this approach, we demonstrate that some of the recent H5 HAs require as few as one or two amino acid mutations to quantitatively switch their receptor preference. Importantly, our approach permitted us to scan the RBS of HAs from various H5 isolates from 2007 for the natural acquisition of necessary RBS features for switch in receptor specificity. Different phylogenetic clades of H5 have important nuances to their RBS structural features, which, in specific instances, dynamically change with the natural sequence evolution. Some features that are present in the parent clade might be lost as this clade diversifies, or this diversification could lead to the addition of features that are critical for human adaptation of H5 HA. Even within a clade, not all sequences have an identical set of structural features. For example, in the case of 2.2.1 HAs, only ~87% have loss of 158 glycosylation. Also, not all 2.2.1 HAs have a deletion in the 130 loop. Data show that the same amino acid mutations that lead to aerosol transmission of Viet04 and Ind05, when introduced in more recent H5 HA, give distinct results (depending on the H5 clade) and, importantly, do not quantitatively switch any of the mutant HAs binding to human receptors. These residues alone cannot be used as reference points to analyze the switch in receptor specificity of currently circulating and evolving H5N1 strains [160]. A question arises as to the relationship between RBS of H5 and H1 HA—its second-nearest human-adapted phylogenetic neighbor in group 1 HAs (Figure 3.3). We have performed a similar structural analysis and have demonstrated that amino acid changes in H5 HA to match features with RBS of pandemic H1N1 (A/South Carolina/1/18) HA led to a quantitative switch in binding to human receptors. However, the current natural evolution of H5 HA has led to acquisition of RBS features along an H2-like path. Human adaptation of H5 HA is one of the key factors involved in the adaptation of the H5N1 virus to the human host for a sustained circulation. Other hallmark factors and genetic signatures such as Glu627/Lys in PB2, PB1-F2 length, activity, and stalk length of neuraminidase have been associated with increased pathogenicity and efficient transmission in humans and ferret animal models. However, it is unclear as to how many of these additional hallmark factors are required for the human adaptation of this subtype. Nevertheless, from a surveillance standpoint, it is critical to investigate amino acid sequence divergence of H5 HA in the context of RBS molecular features in addition to monitoring changes in other viral genes to delineate H5N1 human adaptation.

This work resulted in the following publication:

"Structural determinants for naturally evolving H5N1 hemagglutinin to switch its receptor specificity" published June 06, 2013 in Cell, Volume 153, Issue 7, 1475-1485;
DOI: http://dx.doi.org/10.1016/j.cell.2013.05.035

Abstract

Of the factors governing human-to-human transmission of the highly pathogenic avian-adapted H5N1 virus, the most critical is the acquisition of mutations on the viral hemagglutinin (HA) to "quantitatively switch" its binding from avian to human glycan receptors. Here, we describe a structural framework that outlines a necessary set of H5 HA receptor- binding site (RBS) features required for the H5 HA to quantitatively switch its preference to human receptors. We show here that the same RBS HA mutations that lead to aerosol transmission of A/Vietnam/1203/04 and A/Indonesia/5/05 viruses, when introduced in currently circulating H5N1, do not lead to a quantitative switch in receptor preference. We further demonstrate that HAs from circulating clades require as few as a single base pair mutation to quantitatively switch their binding to human receptors. The mutations identified by this study can be used to monitor the emergence of strains having human- to-human transmission potential.

This work resulted in the following patent:

Human adaptation of H5 influenza, Patent appl. No. PCT/US14/15397, Filed February 7, 2014

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Kannan Tharakaraman, Rahul Raman, Karthik Viswanathan, Akila Jayaraman, Arvind Krishnan, V. Sasisekharan and Ram Sasisekharan for their contributions. This work was funded in part by the National Institutes of Health (R37 GM057073-13) and the National-Research-Foundation-supported Interdisciplinary Research group in Infectious Diseases of SMART.

Chapter 4 : Glycan Receptor binding of the Influenza A Virus H7N9 Hemagglutinin

Summary and Significance

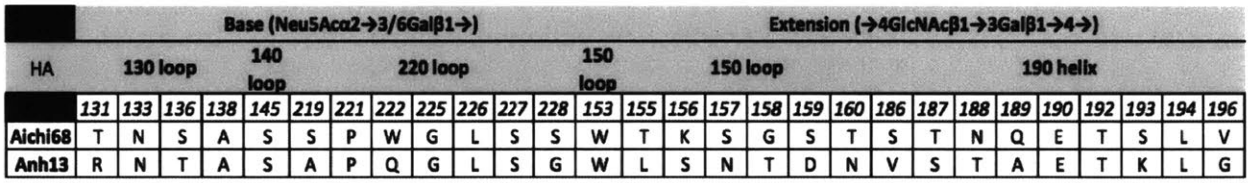
In the previous chapter, I leveraged an improved integrated analysis (implemented in chapter 2) to identify the structural determinants for naturally occurring H5N1 HAs to switch its receptor binding preferences from avian to human glycan receptor binding. This unique integrated approach leveraged bioinformatics analyses, structural modeling, inter-residue network analysis (SIN), and biochemical assays. This approach yielded novel insight into the HA-glycan receptor interaction and such analysis can be extended to study other IAV HA. Leveraging this unique approach and newfound insights, I turned my attention to the HA isolate from the 2013 H7N9 outbreak with the goal of identifying the structural determinants required for H7N9 to increase its binding to human glycan receptors.

In late March of 2013, an H7N9 influenza A virus subtype was identified that could infect humans, causing rapidly progressing severe lower respiratory tract infection and mortality. The advent of H7N9 was of concern for a number of reasons: it could infect humans, the etiology of infection was unclear, and the human population does not have pre-existing immunity to the H7 subtype. Furthermore, earlier sequence analyses of H7N9 hemagglutinin (HA) identified amino acid changes that are known to correlate with improved human receptor binding and impinge on the antigenic properties of the HA. Taken together, these results suggest this virus has pandemic potential. Interestingly, despite the prediction that H7N9 HA should interact strongly with human glycan receptors, it did not show robust human-to-human transmission. Consequently, the goals of this study were to characterize the binding of H7N9 HA to physiologic glycan receptor in human tissues and identify the structural requirements for H7N9 to increase binding to human glycan receptors.

The 2013 outbreak H7N9 HA was expressed recombinantly and showed limited binding to human receptors in the human trachea (as measured by immunofluorescence (IF)). Using the same integrated framework used to enumerate the structural features required for H5N1 to switch its receptor specificity, a single amino acid mutation in H7N9 HA was identified that could cause structural changes within the receptor binding site allowing for enhanced interactions with the human glycan receptor. The mutant H7 showed extensive binding to human glycan receptors present on the apical side of the human trachea, as well as goblet cells. Furthermore, a subset of the H7N9 HA sequences demarcating coevolving amino acids appears to be in the antigenic regions of H7, which, in turn, could impact effectiveness of the current WHO-recommended pre-pandemic H7 vaccines.

Introduction

In late March of 2013, an H7N9 influenza A virus subtype was identified that was found to infect humans, causing rapidly progressing severe lower respiratory tract infection and mortality [164]. As of early May 2013, 131 reported cases have been confirmed, with most cases in mainland China. Detailed sequencing and analysis of the human isolates of H7N9 have offered insights into their potential origin and factors that govern the virus's virulence and pathogenicity [165]. The transmission and pathogenicity associated with this virus is unanticipated for several reasons. First, transmission of H7 viruses from birds to mammals has only been reported rarely [166]. Secondly, until the advent of H7N9, human infections with N9 subtype viruses have not been previously reported. Finally, human infections with other H7 viruses (primarily H7N2, H7N3, and H7N7), even with high-pathogenicity viruses containing a polybasic cleavage site in HA, have primarily resulted in conjunctivitis or uncomplicated illness, with few exceptions [44] [167].



	Base (Neu5Acα2→3/6Galβ1→)														Extension (→4GlcNAcβ1→3Galβ1→4→)													
HA	130 loop				140 loop	220 loop							150 loop		150 loop					190 helix								
	131	133	136	138	145	219	221	222	225	226	227	228	153	155	156	157	158	159	160	186	187	188	189	190	192	193	194	196
Aichi68	T	N	S	A	S	S	P	W	G	L	S	S	W	T	K	S	G	S	T	S	T	N	Q	E	T	S	L	V
Anh13	R	N	T	A	S	A	P	Q	G	L	S	G	W	L	S	N	T	D	N	V	S	T	A	E	T	K	L	G

Figure 4.1 **The key amino acids in the 130, 140, 150, and 220 loops and 190 helix** that bind to the base and extension region of the glycan receptors in Aichi68 and Anh13 are shown. The 150 loop has amino acids (positions 153 and 155) that are involved in contacts with the sialic acid in the base region and those that are involved in binding to extension region.

Analysis of newly arising H7N9 strains, including A/Shanghai/ 1/2013, A/Shanghai/2/2013, and A/Anhui/1/2013, indicates that H7N9 is a reassorted virus incorporating envelope genes from at least two H7 strains (hemagglutinin [HA] from an H7N3 strain and neuraminidase [NA] from an avian-adapted H7N9 strain) with the internal genes from at least two H9N2 avian-adapted influenza strains [165]. Further analysis of the H7N9 gene segments has shown the occurrence of signature amino acids associated with adaptation to human host and virulence (summarized in table 2 in Gao et al. [2013]). H7N9 strains exhibit hallmark mutations that are thought to correlate with increased virulence and potential transmission in animal models such as mice and ferrets, including the E627K mutation in PB2 (which is known to play a key role in human-to-human respiratory droplet transmission [168]). Furthermore, genetic analysis indicates the presence of mutations, in at least some strains, within the M2 ion channel and NA that confer drug resistance to the adamantanes and oseltamivir, respectively.

In contrast to the above analysis, several of the hallmark features found in highly pathogenic influenza strains (e.g., H5N1), including the N66S mutation in PB1-F2, the aforementioned polybasic sequence in the linker between HA1 and HA2, and the PDZ-binding motif in C terminus of NS1, are absent in the H7N9 human isolates analyzed to date. Within this context and given the somewhat confounding genetic and epidemiological evidence of the relative human-adaptation of H7N9, one of the most important outstanding questions is the "status" of the HA protein for these isolates. Characterization of the HA protein is important given its role in virulence [62] and virus neutralization to pre-existing antibodies through antigenic memory [130]. Finally, the receptor-binding properties of HA, governing a given virus's tissue and organismic tropism, are one of the key factors that critically govern aerosol transmissibility, including human-to-human transmission [168].

Previous studies have demonstrated that one of the key properties governing human adaptation of influenza A virus is a "switch" in the glycan receptor-binding specificities of viral HA [169]. Therefore, together with hallmark mutations in other genes, such as PB2, describing mutations in HA that lead to such a "switch" becomes important for surveillance purposes. From the standpoint of human tissue tropism, the

HA from human-adapted viruses, including pandemic strains, shows extensive binding to the apical surface of human upper respiratory tissues (such as trachea) and also shows characteristic binding to mucin-secreting nonciliated goblet cells on the apical surface and to submucosal glands in ferret respiratory tract [162] [170] [41]. Through lectin staining, it has been demonstrated previously that these regions in human tracheal sections pre- dominantly display diverse glycan receptors terminated by $\alpha 2 \rightarrow 6$ sialic acid linkage (human receptors) [41] [139]. We have previously demonstrated with H1, H2, and H3 subtypes that this binding property of human-adapted viruses is one of the key factors that correlates with their ability to efficiently transmit via respiratory droplets in ferrets—a well-established animal model to measure the potential for airborne human-to-human transmission [154] [155].

In the case of H7N9, based on the presence of a leucine residue in the 226 position (H3 numbering) of the HA, earlier studies have predicted that this HA would have strong binding to human receptors [164] [171]. These recent studies note that HA-glycan receptor interaction is a critical property of the virus and that experimental characterization of this property for the H7N9 subtype is important.

Here, we report the molecular and structural features of the glycan receptor-binding site (RBS) of H7N9 HA and the experimental characterization of its binding to physiological glycan receptors in the human respiratory tract. Contrary to the predicted strong binding to human receptors, H7N9 HA shows limited binding to these receptors in the human upper respiratory tract when compared to human receptor binding of other human-adapted HAs. The experimentally observed limited binding to human receptors by H7N9 HA is consistent with the analysis of its RBS structural features. The structural and sequence analyses further point to a single Gly228/Ser amino acid change, which modifies the network of inter-residue contacts in the RBS for more optimal contacts with both avian and human receptors. Consequently, introducing this amino acid change in H7N9 HA RBS resulted in a mutant HA that extensively bound to the apical surface of human tracheal tissue sections in a fashion similar to that of other human-adapted HAs. Our findings therefore provide important insights into the physiological glycan receptor tropism of the current H7N9 HA. We also report an increase in human receptor binding should a single Gly228/Ser amino acid change occur in this HA in the context of monitoring the evolution of this emerging subtype, especially as it continues to circulate in humans. Finally, we report amino acid substitutions in the H7N9 HA sequences that distinguish the evolution (including antigenic sites) of this recently emerged sub- type from past H7 isolates, which, in turn, has implications for vaccine development strategies.

Experimental Methods

Cloning, Baculovirus Synthesis, and Mammalian Expression and Purification of HA

Anh13 WT and G228/S mutant HA sequences were codon optimized for mammalian expression, synthesized (DNA2.0, Menlo Park, CA), and subcloned into modified pcDNA3.3 vector for expression under CMV promoter. Recombinant expression of HA was carried out in HEK293-F Free Style suspension cells (Invitrogen, Carlsbad, CA) cultured in 293-F Free Style Expression Medium (Invitrogen, Carlsbad, CA) maintained at 37 degrees Celsius, 80% humidity, and 8% CO2. Cells were transfected with Polyethyleneimine Max (PEI-MAX, PolySciences, Warrington, PA) with the HA plasmid and were harvested 7 days post-infection. The supernatant was collected by centrifugation, filtered through a 0.45 mm filter system (Nalgene, Rochester, NY), and supplemented with 1: 1,000 diluted protease inhibitor cocktail (Calbiochem filtration and supplemented with 1:1,000 diluted protease inhibitor cocktail [EMD Millipore, Billerica, MA]). HA was purified from the supernatant using His-trap columns (GE Healthcare) on an AKTA Purifier FPLC system. Eluting fractions containing HA were pooled, concentrated, and buffer exchanged into 13 PBS (pH 7.4) using 100 kDa MWCO spin columns (Millipore). The purified protein was quantified using the BCA method (Pierce, Rockford, IL).

Homology Modeling of HA

A structural model of Anh13 HA was built using the MODELER homology modeling software. The crystal structure of A/Netherlands/219/2003 (Neth03) HA (PDB:

4DJ6) was used as a template to build the model. The structural model of Anh13 bound to avian receptor was constructed by superimposing the HA1 from co-crystal structure of Neth03-avian receptor complex (PDB ID: 4DJ7) on Anh13 HA1. The structural model of Anh13 in complex with human receptor was constructed by superimposing the HA1 from co-crystal structure of Aichi68-human receptor complex (PDB ID: 2YPG) with HA1 of Anh13. The final models were subject to energy minimization (500 steps conjugate gradient + 500 steps steepest descent) with potentials assigned using AMBER force field.

Coevolution, Phylogeny, and Selection Analyses of H7 HA Sequences

A total of 625 non-redundant full-length H7 HA sequences were downloaded from GISAID. To further eliminate redundancy, the sequences were grouped according to subtype, host, year, and country, and a representative sequence was chosen from each group. This led to a total of 231 HA sequences. Coevolving groups of amino acids were predicted using the CAPS online server for protein coevolution (http://bioinf.gen.tcd.ie/caps/). The results indicate functionally or structurally linked regions that are subjected to strong selective constraints. A phylogeny tree was constructed from the 231 HA sequences using the neighbor-joining method found in MEGA 5.1 software (http://www.megasoftware.net/). Protein-coding nucleotide sequences were extracted for the 114 Eurasian HA sequences, and the region-encoding residues 50–230 of HA1 were employed for finding individual codons under diversifying/positive selection. Positively selected sites were predicted using DataMonkey (http://www.datamonkey.org/), which uses a normalized dN-dS > 0 at p value < 0.1 threshold to detect positive selection.

Binding of HA to Human Tissue Sections

The human tracheal epithelia has been extensively benchmarked as a tissue section representative of the human upper respiratory tract that is a predominant physiological target site for human-adapted influenza A viruses [151] [172] [161] [163]. The apical surface and submucosal regions of the human trachea have been shown to predominantly display human receptors [173] [172] [163]. On the other hand, human alveolar tissue sections representative of deep lung region have been shown to predominantly express avian receptors and are typically stained by HA from avian-adapted influenza A viruses [173] [174]. Paraffinized human tracheal and alveolar (US BioChain) tissue sections were deparaffinized, rehydrated, and incubated with 1% BSA in PBS for 30 min to prevent nonspecific binding. HA was precomplexed with primary antibody (mouse anti-63His tag, Abcam) and secondary antibody (Alexa fluor 488 goat anti-mouse, Invitrogen) in a molar ratio of 4:2:1, respectively, for 20 min on ice. The tissue binding was performed over two different HA concentrations (40 mg/ml and 20 mg/ml) by diluting the precomplexed stock HA in 1% BSA-PBS. Tissue sections were then incubated with the HA-antibody complexes for 3 hr at room temperature (RT). The tissue sections were counter- stained by propidium iodide (Invitrogen; 1,100 in TBST). The tissue sections were mounted and then viewed under a confocal microscope (Zeiss LSM 700 laser scanning confocal microscopy). Sialic-acid-specific binding of HAs to tissue sections was confirmed by loss of staining after pretreatment with sialidase A (from *Arthrobacter ureafaciens*, Prozyme). This enzyme has been demonstrated to cleave the terminal Neu5Ac from both Neu5Aca2/3Gal and Neu5Aca2/6Gal motifs. In the case of sialidase pretreatment, tissue sections were incubated with 0.2 units of sialidase A for 3 hr at 37 degrees Celsius prior to incubation with the proteins. Pretreatment of human tissue sections with sialidase A resulted in complete loss of HA staining.

Capturing Network Inter-amino Acid Contacts for RBS Residues

The coordinates of Neth03 H7 HA-avian receptor (PDB ID: 4DJ7) and Aichi68 H3 HA-human receptor complexes (PDB ID: 2YPG) were uploaded into the PDBePISA server (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) to determine key residues in the HA RBS that make contact with the corresponding glycan receptor (interface cutoff of 30% was used). For these residues, their environment was defined using a distance threshold of 7 A° , and the contacts, including putative hydrogen bonds (including water-bridged ones), disulfide bonds, pi bonds, polar interactions, salt bridges, and Van der

Waals interactions (nonhydrogen) occurring between pairs of residues within this threshold distance, were computed as described previously [157]. These data were assembled into an array of eight atomic interaction matrices. A weighted sum of the eight atomic interaction matrices were then computed to produce a single matrix that accounts for the strength of atomic interaction between residue pairs within the RBS, using weights derived from relative atomic interaction energies [157]. The interresidue interaction network calculated in this fashion generates a matrix that describes all the contacts made by critical RBS residues with spatial proximal neighboring residues in their environment. For each element i, j is the sum of the path scores of all paths between residues i and j. The degree of networking score for each residue was computed by summing across the rows of the matrix, which was meant to correspond to the extent of "networking" for each residue. The degree of networking score was normalized (RBSN score) with the maximum score for each protein so that the scores varied from 0 (absence of any network) to 1 (most networked).

Results

Given that H7N9 is a newly emerged subtype, there is a limited set of HA sequences available for this subtype to do a comprehensive analysis of sequence evolution of its RBS. Therefore, in order to understand glycan receptor-binding properties of H7N9 HA, we chose a representative human isolate A/Anhui/1/2013 (Anh13). Because there is no crystal structure available, we constructed a homology-based structural model of Anh13 HA and compared its RBS with H3 HA (its phylogenetically closest human-adapted HA [Figure 4.2]). Anh13

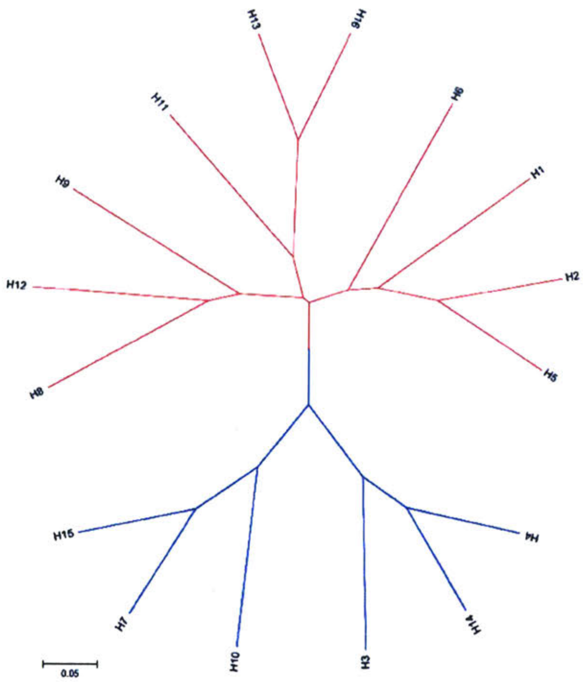


Figure 4.2 **Phylogenetic Distance between Different HA Subtypes**. Branches leading to group 1 & 2 HAs are labeled and colored in red and blue, respectively. Closely related subtypes are located on branches close to one another.

was chosen because it shares substantial sequence identity with many other reported strains of H7N9, including A/Shanghai/2/ 2013 and A/Hangzhou/1/2013. Additionally, as noted previously[164], although there are changes in the HA sequence of Anh13 compared to the HA sequence of other reported H7N9 strains, including A/Shanghai/2/2013, the HA of Anh13 contains a key Q226L mutation that has been reported to be important for altered receptor specificity for group 2 viruses, including H3 and H7. As such, Anh13 (and related viruses) are more likely than viruses such as A/Shanghai/2/ 2013 to be of concern from the standpoint of altered receptor specificity and, hence, human transmissibility.

Earlier, we defined a framework that incorporated descriptors of the structural topology of the human glycan receptor as well as inter-amino acid interaction networks within the RBS to define the molecular features of the RBS of H5 HA for high-avidity/specificity binding to human glycan receptors[173] [157]. We therefore compared the molecular features of the Anh13 RBS with those of H3 HA (from A/Aichi/1/68 or Aich68, a strain from the 1967–1968 pandemic), which was recently co-crystallized with both avian and human receptors[144]. Also of importance, because Aich68 represents a human-adapted virus, comparison of its HA to that of Anh13 provides an important benchmark to address the question of whether the HA from Anh13 shares structural characteristics with HAs of human-adapted viruses and, if not, which structural characteristics are missing. Structural analyses of both H7 and H3 HAs indicate that the structural topology of the human glycan receptor bound in the RBS of HA is such that it has a clearly defined base region consisting of the terminal Neu5Acα2→6Galβ1→ motif and an extension region consisting of at least a disaccharide →4GlcNAcβ1→3Galβ1→. On the other hand, the topology of the avian receptor bound within the HA RBS is such that majority of the contacts with residues within the RBS involve the terminal Neu5Acα2→3Galβ1→motif in the base region (Figure 4.3). The 130 (residues 131–138), 140 (residues 140–145) and 220 (residues 219–228) loops in the RBS make critical contacts with the disaccharide motif in the base region and thus play a key role in dictating the avian or human receptor-binding preference (residue

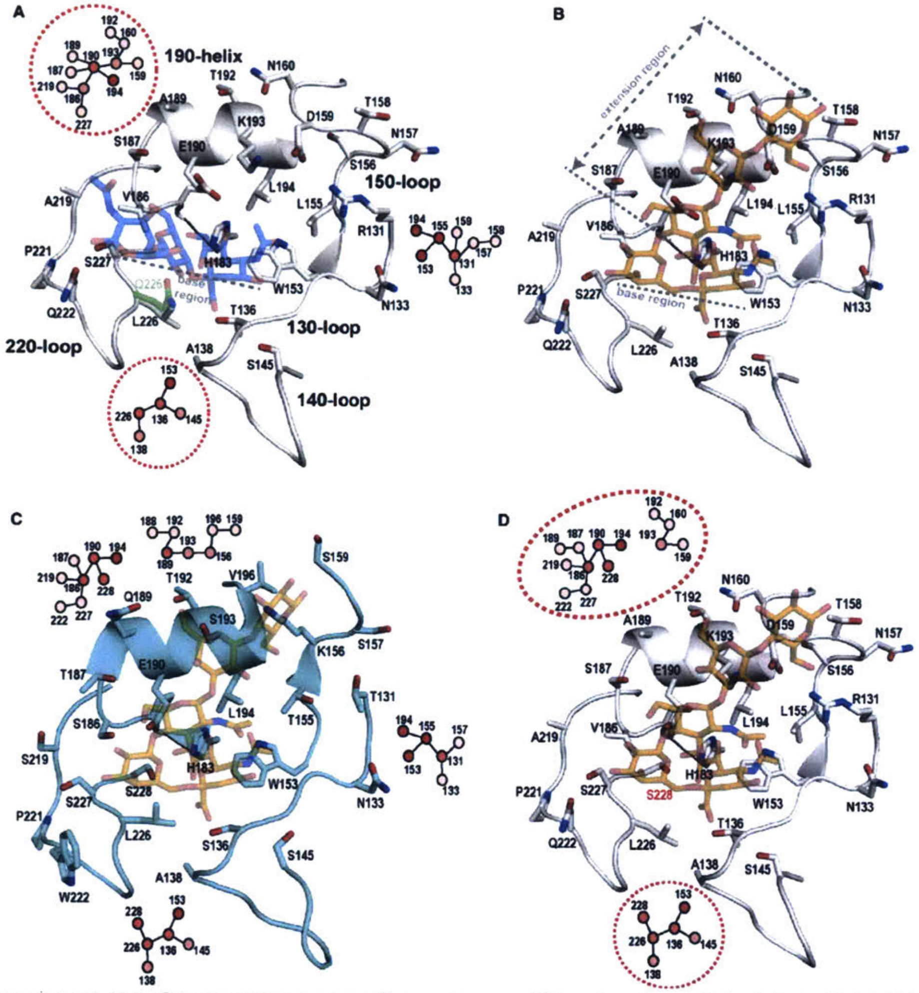


Figure 4.3 **Structural HA-Glycan Receptor Complexes**. Structural model of Anh13 H7 HA RBS in complex with avian receptor. The avian receptor is shown as a stick at 40% transparency with carbon atom colored in blue. The loop and helix regions used to define molecular features of the HA RBS are shown. The side chains of the key amino acids in these regions that make contact with the glycan receptor are shown. The side chain of Q at the 226 position observed in all other H7 HAs (prior to H7N9 outbreak) is shown as stick with 40% transparency (carbon atom colored green). The networks of inter-residue interaction contacts are shown as two-dimensional maps near the corresponding residue positions. The map is composed of circular nodes representing the amino acid positions (which are labeled above the nodes) and is colored according to the degree of inter-residue contact (lightest shade of red indicating lowest contact and darkest shade of red indicating highest contact). (B) The structural model of Anh13 H7 HA RBS in complex with human receptor (shown as a stick with 40% transparency and carbon atom colored in orange). (C) Aichi68 H3 HA RBS-human receptor complex as observed in the X-ray cocrystal structure (PDB ID: 2YPG). The inter-residue interaction network of the key RBS residues is similar to what was shown in (A). Note the similarities in the network of residues in the 130, 140, and 220 loop between the H3 and H7 HA (expanded in Figure 4.1). On the other hand, the network involving residues in the 190 helix is quite different, and this difference is brought about by the amino acid differences and also by the Gly in H7 HA versus Ser in H3 HA in the 228 position. (D) The structural model of the G228→S mutant of Anh13 H7 HA RBS in complex with human receptor. Note that the network involving residues in the 190 helix in the mutant is more similar to that observed in H3 HA than the WT. The inter-residue contact networks that are different between the mutant and WT HA are shown in red dotted circle

position numbering is based on Aichi68 HA-glycan co-crystal structure [Protein Data Bank ID: 2YPG]). Additionally, residues within the 190 helix (residues 190–196) and apart of the 150 loop (residues 156–160), if properly positioned, make critical contacts with the extension region of the human receptor. Taken together, these five loops and one helix in the RBS that make contact with the base and potentially with the extension region of the human receptor and their network of interactions with spatially proximal residues in the RBS constitute a complete set of molecular features that should be analyzed to understand the receptor-binding preference of an HA. Comparison of these features between the Anh13 and Aichi68 HA shows many similarities, as well as some important differences (Figure 4.1). First, many of the residues in the 130 loop, 140 loop, and 220 loop are similar between the HA of Anh13 and Aichi68. Based on this level of similarity, we find that the network of inter-residue contacts involving these residues is also similar (Figure 4.3A and C). The key structural differences noted in this analysis include part of the 190 helix and the 150 loop required for contact with the extension region of the human receptor. In the Aich68 HA, the

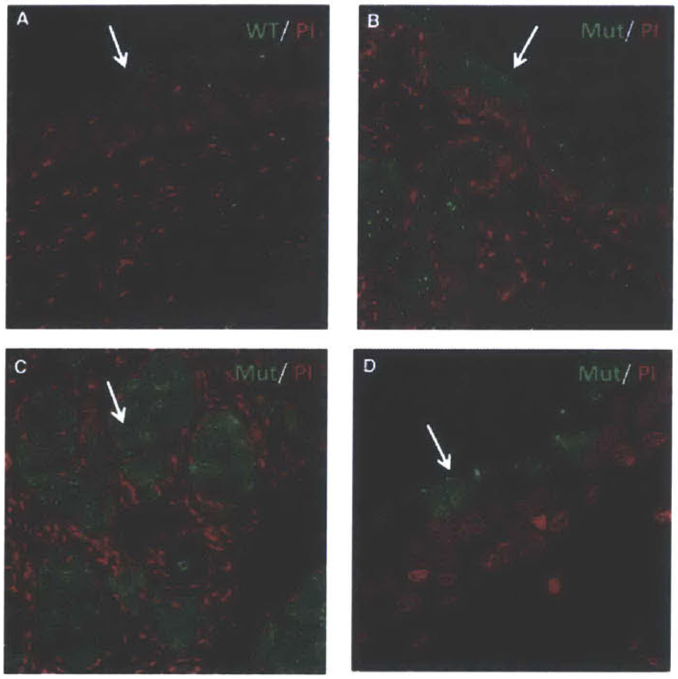


Figure 4.4 **Staining of Human Trachea with WT and G228S** A/Anhui/1/13 Hemagglutinin. (A–D) Human paraffinized tissue sections were stained with recombinant HAs expressed and purified from 293 F cells. Specific staining by recombinant HA (in green) is also demarked by white arrows. The recombinant HAs were pre-complexed with primary anti-His and Alexa fluor 488 tagged (green) secondary antibodies (for multivalent presentation) before adding to the tissue sections. The WT protein did not stain the trachea (A) as intensely as the G228S mutant HA (B). The G228S HA showed intense staining of the apical surface of the trachea (marked by white arrow). One key feature of the G228S protein is the staining of the submucosal gland (C) and the goblet cells (D) in the human trachea. The staining to goblet cells is similar to staining by other human-adapted influenza A virus HA. The tissue was counterstained with propidium iodide (PI) shown in red. Images in (A), (B), and (C) were captured at 25× magnification and the image in (D) was captured at 63× magnification.

residues interacting with the extension region include Q189, S193, K156, and S159 (Figure 4.3C), whereas, in H7 HA, these residues include K193, T158, D159, and N160 (Figure4.3B). Furthermore, in the case of the H7 HA, R131 is positioned to make an additional contact with the extension region (Figure 4.3B). In addition to the differences in the amino acids at these positions—and also of note—are differences between Aich68 and Anh13 in the interresidue interaction network governed by the residue at position 228. This position is a Ser in Aichi68 but is a Gly in Anh13. The S228 position in H3 is critical for the interamino acid network involving S186, T187, and E190, which positions E190 to make critical contacts with the sialic acid of both avian and human receptors (4.3C). On the other hand, G228 in H7 HA does not possess this interamino acid network and, therefore, the network of interresidue contacts involving E190 in H7 is different from that of H3 HA and instead includes 193 (and its network) and 189 (4.3A).

Extending this analysis, the E190 in H7 HA is positioned to make additional contacts with the extension region instead of the critical contact with the sialic acid in the base region (Figure 4.3B). Therefore, Anh13 H7 HA lacks at least two critical contacts involving the 190 and 226 HA positions with the Neu5Acα2→3Galβ1→terminal motif as observed in avian- adapted HAs. In the case of human receptor contacts, Anh13 has lower contacts with the base region owing to the absence of the S228 residue. Our structural analysis pointed to the H7N9 HA having a substantially lower binding to avian receptors than typical avian-adapted HAs, as well as lower binding to human receptors than the human-adapted H3 HAs. Furthermore, the RBS of H7N9 HA is such that a single G228/S amino acid change would modify the interamino acid network in the RBS to position the E190 and S228 residues for optimal contacts with both avian and human receptors.

To validate the structural analyses, we analyzed binding of Anh13 to tissue sections, representative of the human respiratory tract, which display the physiologically relevant receptors for influenza A viruses (Figure 4.5A). Anh13 stained the apical surface and submucosal region of the human trachea, which express glycans known to be receptors of human-adapted viruses[172] [163]. However, the intensity and the extent of tracheal apical surface staining by Anh13 were substantially lower than what is typically observed for human-adapted HAs (Figure 4.5A). Furthermore, Anh13 HA also showed minimal binding to deep-lung alveolar section—a region that is extensively stained by avian-adapted HAs (Figure 4.5B). In contrast, the G228/S mutant HA showed a dramatic and significant increase in the extent and intensity of staining to apical surface of the tracheal section, including extensive staining of nonciliated goblet cells in a fashion similar to that of other human-adapted HAs and Sambucus nigra agglutinin I (SNA I) (Figure 4.4 and Figure 4.5C). Interestingly, the G228/S mutation also substantially increased its binding to the alveolar sections (Figure 4.5B). These results are consistent with the structural analyses of the RBS features of H7N9 HA.

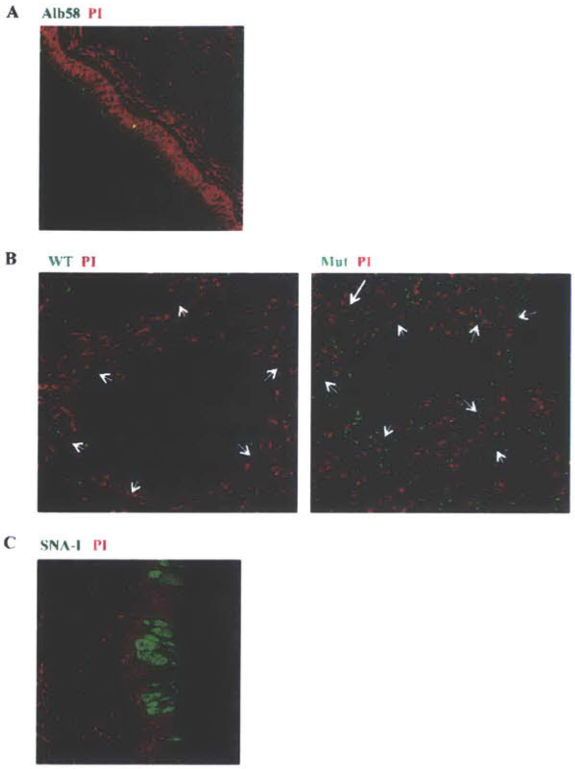


Figure 4.5 **Binding of HAs and SNA-I to Human Respiratory Tissue Sections**, Related to figure 4.4 (A) A/Albany/6/58 (Alb58) pandemic H2N2 HA shows extensive binding to apical surface of tracheal tissue section (predominantly expresses human receptors) even at HA concentration of 10 µg/ml. (B) Staining of paraffinized human alveolar section with WT and G228S A/Anhui/1/13 hemagglutinin (HA). The WT protein did not stain the alveolus as intensely as the G228S mutant HA. The staining of alveolar section by WT and mutant HA is also indicated using white arrows. (C) Staining of paraffinized human tracheal section with Sambucus nigra agglutinin I (SNA I), a lectin known to specifically bind to human receptors. SNA-I showed staining of the apical surface of the human trachea. SNA-I also stained the mucin secreting goblet cells similar to the Anhui 13 G228S mutant HA. Image was captured at 40X magnification. All tissue sections were counterstained with propidium iodide (PI) shown in red. The staining of HA and SNA I is shown in green.

Finally, in the context of H7N9 evolution, two mutations (174S and 226L) appear to be unique to the H7N9 HA sequences. As noted above, the residue at 226 is a critical determinant of the receptor-binding specificity of H7 HA, with human viruses favoring L/I and avian viruses favoring Q. These two positions are also part of a larger cluster of coevolving positions (122A, 174S, 186V, 202V, and 226L), all within the 50–230 HA region, which demarcates the virus from its previous H7 ancestors (see the experimental methods). The above observations indicate that the H7N9 HA has evolved to be distinct from its predecessors. Previous H7 strains carrying single mutations (from the coevolving cluster) are predominantly from the Eurasian lineage, suggesting that viruses from this lineage have higher potential to generate variants when compared to the American lineage (Figure 4.6). Nucleotide analyses of the RBS-proximal region of HA (residues 50–230 of HA1) of Eurasian sequences from 1902 to 2013 show strong diversifying (positive) selection at 156 (see the methods section). The same position has been shown to be under selection pressure in the H1 subtype as well[175].

Previously, we defined a quantitative metric to compare the antigenicity of two HAs. Briefly, the metric, called antigenic intactness (AI), is directly proportional to the fraction of residues conserved in the immunodominant antigenic sites between two HAs. In our previous work, we showed good agreement between AI values and antigenic relatedness metric computed from ferret antisera HA inhibition (HI) cross-reactivity data[176], indicating that AI values could be applied to predict vaccine-induced cross-reactive antibody responses. Critically, strains that are antigenically related had AI > 80%, whereas strains that are not related to each other had AI < 80%[176]. Keeping 80% as the cutoff, the AI values between the recent World Health Organization (WHO)-recommended H7 vaccines strains (A/Canada/rv444/ 2004 [H7N3], A/mallard/Netherlands/12/2000 [H7N3], and A/ New York/107/2003 [H7N2]) and the A/Anhui/1/2013 H7N9 HA were computed. The H7N9 HA has AI values of 67%, 89%, and 70% with A/Canada/rv444/2004 (H7N3), A/mallard/Netherlands/12/2000 (H7N3), and A/New York/107/2003 (H7N2), respectively, suggesting that only /mallard/Netherlands/ 12/2000 (H7N3) may be effective as a vaccine component. However, amino acid differences between A/Anhui/1/2013 (H7N9) and A/mallard/Netherlands/12/ 2000 (H7N3) at key antigenic sites (122 in site A and 188 and 189 in site B) might limit this mallard strain to be effective.

In the context of the Q228S mutation and given that H7N9 is a subtype that has just emerged in the human population, it is difficult to predict whether this mutation alters antibody response to this region in this subtype. However, it is generally known that the 220 loop, which includes the 228 position, is an antigenic site for some HA subtypes (224 and 225 are part of antigenic site Ca in H1 subtype; 222 is part of antigenic site I-C in H2 subtype; and 220 is part of antigenic site D in H3, a subtype that is phylogenetically closest to H7); thus, antibody response targeting this region in H7N9 cannot be excluded. As such, it will be important to complete serological analysis to confirm the vaccine implications of our AI analyses.

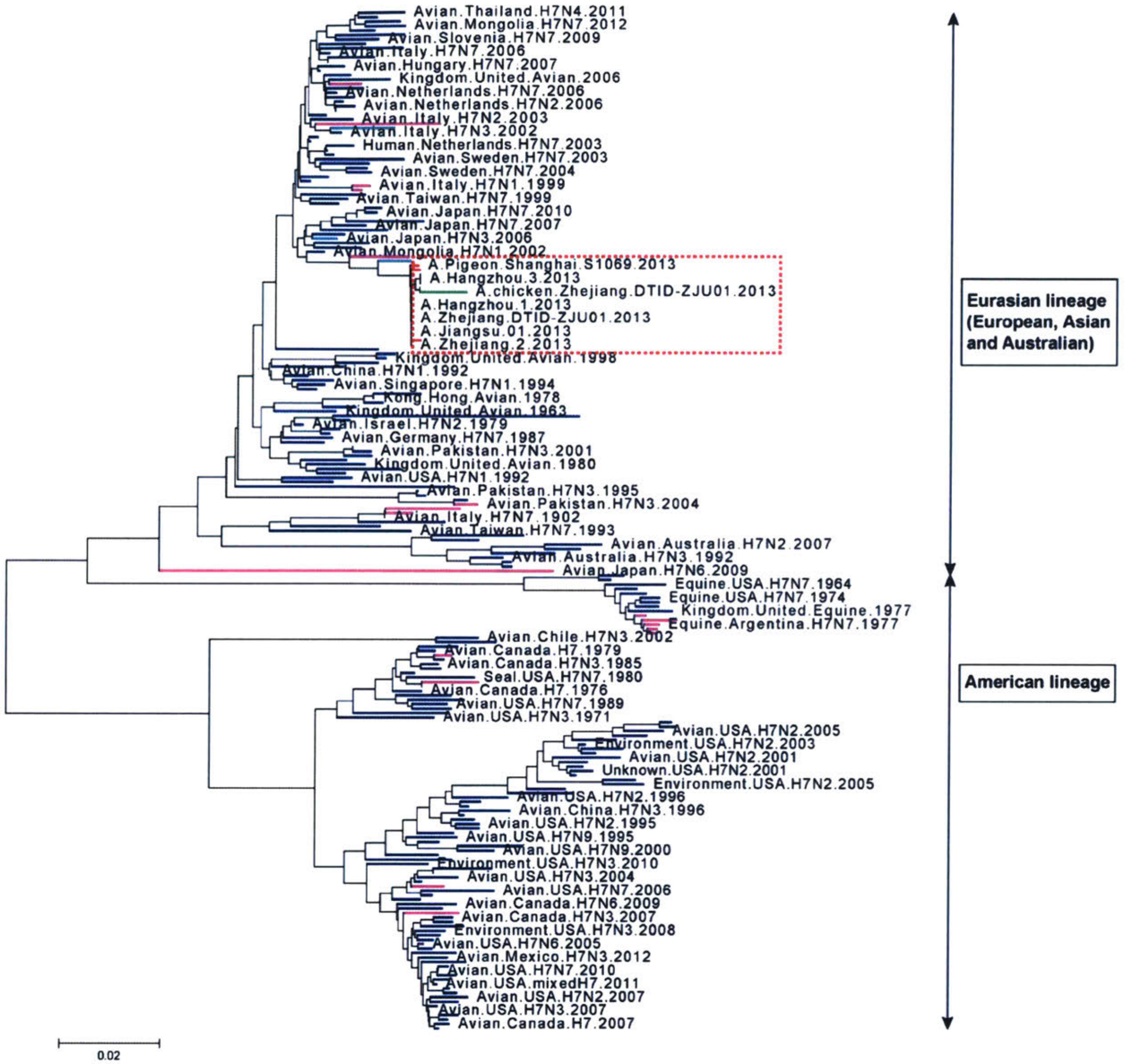


Figure 4.6 **Phylogeny Tree of H7 HA Amino Acid Sequences**. The tree was constructed using 231 full-length, non-redundant amino acid sequences using the Neighbor-Joining method found in the MEGA 5.1 software. The two major H7 lineages are labeled on the right hand side by double-sided arrows. Branches are color-coded according to the number of coevolving residues found in the H7 sequence (red: 5; blue: 4; green: 2; magenta: 1; black: 0). A red rectangular box marks occurrence of the novel H7N9 virus on the tree.

Conclusion

The emergence of an H7N9 influenza A virus subtype poses a significant global health concern, given that it has led to severe infection and mortality in humans. Although preliminary genetic analysis of this subtype has led to predictions about its human host adaptation based on hallmark genetic signatures, including strong binding to human receptors, this virus has not yet caused widespread infection in humans resulting from aerosol human- to-human transmission. Earlier studies have highlighted the importance of determining the glycan receptor-binding property of H7N9 HA because this property is one of the many key factors that govern human adaptation of the virus [171]. Therefore, in this study, we experimentally characterized glycan receptor-binding properties of H7N9 HA to assess the human adaptation of this HA. To the best of our knowledge, this is the first report on the glycan binding data of the H7N9 HA.

Our structural analyses and binding of H7N9 HA to human respiratory tissues demonstrate that it possesses a distinct binding tropism when compared to either an avian-adapted or a human- adapted HA. Our findings shed light on the distinct tropism of this H7N9 HA that is contrary to the expected strong human receptor-binding preference predicted from earlier sequence analyses [164] [171]. This distinct tropism would likely impinge on the aerosol transmissibility of the H7N9 viruses in ferrets when compared to the efficient transmission observed in the past pandemic viruses. The limited human and avian receptor binding of the H7N9 HA raises a question as to whether this currently circulating subtype is an intermediate in the adaptation to the human host. Answering this question at this point in time is limited by the availability of sequence information for this recently emerged subtype. Nevertheless, our structural and experimental analyses point to an important role for G228/S amino acid change in the RBS, should it emerge in the current H7N9 HA, in substantially increasing its binding to human receptors in the human respiratory tract. Of note—within the context of this study—is the increased binding of the mutant HA to the noncilicated goblet cells in the human trachea. This is significant because binding to goblet cells is one of the hallmarks of human-adapted HAs[172] [162] [41]. Even in cultures of differentiated human airway epithelial cells, the human influenza A viruses are found to predominantly infect noncilicated cells as compared to avian influenza A viruses, which target the cilicated cells [162].

Previously, several groups, including us, have studied the receptor specificity of influenza HA using “prototypic” glycan arrays containing limited sets of glycans capped with a2–3 or a2–6 linked sialic acid. We find that, in most cases with H1, H2, and H3

HA, the observed binding on glycan array correlated to observed binding of these HAs on physiological glycans from respiratory tract tissue sections. However, as noted in our previous study of HA from H7N2 A/Netherlands/219/2003, we observed that introduction of 226L and 228S resulted in significant staining of the apical region of tracheal tissues, despite having modest binding to a2–6 sialylated glycans on the glycan array [177]. In the present study, with Anh13, we again observed that the presence of 226L and 228S on the HA results in significant binding to the apical and submucosal regions of the trachea that was not captured by the affinity of the HA to limited set of glycans in the array (data not shown). This result warrants an investigation into binding to a more diverse set of a2–6 sialylated glycans, particularly for H7HAs, and further highlights the need for caution in interpreting prototypic glycan array data during surveillance of H7N9 viral evolution.

The significance of the change in residue 156 in receptor binding or other HA function is unclear, although the neighboring 158 glycosylation is known to have an influence on human receptor binding[147]. Significantly, a subset of the H7N9-demarcating coevolving positions (122, 186, and 202) appears to be in the antigenic regions of H7, which could have implications on the effectiveness of the current WHO-recommended prepandemic H7 vaccines. Furthermore, the reported poor immunogenicity of vaccine candidates based on H7N9 is a key challenge for potential vaccine strategy. The fact that the wild-type (WT) virus binds poorly to human receptor supports the notion of poor uptake by human cells to engender an appropriate human immune response. Mutant forms of H7N9 HA, such as G228S with higher specificity to human receptors, can potentially have important applications for the generation of appropriate vaccine countermeasures. Of course, a critical component of vaccine assessment is the use of serological assays to investigate cross-reactivity of heterologous strains. In this context, we have previously demonstrated the correlation between AI score with cross-neutralization responses based on WHO data [176]. Taken together, the mutations on the antigenic regions of H7N9 HA, together with the results of the AI analysis, could impinge on H7 vaccine development.

In summary, our study reports on the experimental glycan-binding properties of the H7N9 HA and also reports on the effects of a single amino acid G228/S change in dramatically increasing glycan receptor binding of this HA. In light of the continued circulation of H7N9 in human subtypes, our study facilitates monitoring the evolution of H7N9, including the acquisition of amino acid changes such as G228/S that would make it closer to human adaptation. Our study warrants further investigation of introducing this single amino acid change to H7N9 virus in the context of other genetic changes characteristic of human-adapted viruses (such as K627 in PB2) and studying its transmission properties in ferrets. This analysis therefore sets the stage for future in vitro studies in human respiratory tract cell cultures and in vivo studies in ferret and mice to investigate the replication potential, virulence, pathogenicity, and respiratory droplet transmission using these recombinant WT and mutant HA H7N9 viruses. Taken together, these findings have important implications for surveillance of H7N9 mutations in clinical settings, as well as for vaccine development efforts.

Abstract

This work resulted in the following publication:

"Glycan receptor binding of the influenza A virus H7N9 hemagglutinin." published June 06, 2013 in Cell, Volume 153, Issue 7, 1486-1493;

DOI: http://dx.doi.org/10.1016/j.cell.2013.05.035

Abstract

The advent of H7N9 in early 2013 is of concern for a number of reasons, including its capability to infect humans, the lack of clarity in the etiology of infection, and because the human population does not have pre-existing immunity to the H7 subtype. Earlier sequence analyses of H7N9 hemagglutinin (HA) point to amino acid changes that predicted human receptor binding and impinge on the antigenic characteristics of the HA. Here, we report that the H7N9 HA shows limited binding to human receptors; however, should a single amino acid mutation occur, this would result in structural changes within the receptor binding site that allow for extensive binding to human receptors present in the upper respiratory tract. Furthermore, a subset of the H7N9 HA sequences demarcating coevolving amino acids appears to be in the antigenic regions of H7, which, in turn, could impact effectiveness of the current WHO-recommended prepandemic H7 vaccines.

This work resulted in the following patent:

Human adaptation of H7 influenza, Patent appl. No. PCT/US14/896062, Filed December 4, 2014

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Kannan Tharakaraman, Rahul Raman, Karthik Viswanathan, Akila Jayaraman, David Johnson, Zach Shriver, V. Sasisekharan and Ram Sasisekharan for their contributions. This work was funded in part by the National Institutes of Health (R37 GM057073-13) and the National-Research-Foundation-supported Interdisciplinary Research group in Infectious Diseases of SMART.

Chapter 5 : Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic

Summary and Significance

Avian-adapted viruses, like H5N1, continue to pose a significant pandemic threat. Due to its antigenic novelty to human hosts, host switching and efficient respiratory droplet transmission are the rate limiting steps in the emergence of an H5N1 pandemic. Consequently, Chapter 3 focused on identifying the structural determinants for H5N1 to switch its receptor binding specificity from avian-to-human. These studies of the HA-glycan receptor interaction provided a framework for improved pandemic surveillance, and ultimately could help identify IAV strains with high pandemic potential circulating in non-human hosts.

However, H5N1 is not the only virus that poses a pandemic threat. Viruses, such as H3N2, that circulate seasonally in people also pose a threat, owing to the fact that H3 subtypes circulate and evolve in both human and non-human hosts such as swine and birds. If an avian- or swine-adapted H3N2 (with significant antigenic divergence from circulating human-adapted H3N2 HAs) gained the ability to infect humans, it could give rise to another H3N2 pandemic. Thus, devising a pandemic risk assessment strategy for HAs derived from H3N2 circulating in non-human hosts requires the investigation of both glycan binding properties and antigenic properties.

In this chapter, leveraging components of our earlier integrated approach, I focus on studying H3N2's HA's antigenic properties and glycan receptor binding properties, with the goal of assessing the likelihood of re-emergence of a pandemic strain from a non-human host. Specifically, this study focuses on integrating bioinformatics tools and experimental methods capable of measuring antigenic properties into the pandemic risk assessment. Using this approach, several H3 strains circulating in swine and birds were identified that possess hallmark features of viruses that could re-emerge and give rise to another H3N2 pandemic in humans (i.e. a high degree of antigenic similarity to the 1968 pandemic H3 and human glycan receptor binding). Furthermore, sera taken from rabbits immunized with recent seasonal H3 vaccine strains failed to recognize these potentially pandemic HAs. These results indicate a severe risk for an emerging pandemic. Consequently, measures should be taken to update the seasonal H3N2 vaccine to include strains that could provide protection against these potentially re-emerging H3s.

Introduction

Influenza A viruses pose a major public health problem, causing seasonal epidemics and occasional—but devastating—global pandemics [33] which negatively impact the global economy. Until recently, influenza pandemics were thought to be associated with the introduction of new HA subtypes into the human population [178] [179] [180] [181]. Indeed, two of the twentieth century pandemics – the 1957–58 H2N2 Asian Flu and the 1967–68 H3N2 Hong Kong Flu - introduced new HA subtypes into the human population[180] [181]. The surface glycoprotein HA of the influenza A virus is the main target of the immune system and mutations on the globular head region (residues 50–230 of HA1, H3 HA numbering used) of this protein determine antigenic novelty, species adaptation, and transmission[182].

Birds are natural reservoirs for influenza A viruses and avian-adapted viruses either directly crossover to humans (through direct contact) or do so with the help of intermediate swine species. Influenza A viruses rapidly evolve (through antigenic drift) in humans as a consequence of both the complex response of human immune system and rapid geographical movement of human population. In contrast to their rapid antigenic evolution in human hosts, the antigenic evolution of influenza A viruses in avian and swine occurs at a much slower rate[183] [184] [185] [186]. As a consequence of these factors, the human immunity to past pandemic strains fades over time, thus enabling antigenically “intact” viruses in avian and swine species to reemerge and begin a new infection cycle in humans. For example, although H2N2 subtype does not currently circulate in the human population, viruses carrying HA that are antigenically similar to the 1957–58 pandemic H2N2 virus continue to circulate in avian species[187]. Among the subtypes that continue to circulate in humans (H1N1 and H3N2), the 2009 H1N1 outbreak offers a practical example of how HA from a swine strain that is antigenically similar to 1918 pandemic H1N1 HA can be reintroduced into the human population[52]. The question remains of whether this trend is observed in H3N2, given that there has been a high rate of antigenic drift in human H3 subtype[188] [54] [189] since the emergence of 1968 pandemic H3N2. Critically, average hospitalizations and mortality rates were found higher for seasons dominated by A/H3N2 viruses compared to seasons dominated by influenza B or A/H1N1 [190] [191] [192].

The H3N2 pandemic began in 1968 and was caused by a human-adapted H2N2 virus that obtained avian H3 and PB1 genes through reassortment [181]. The HA of both 1957 and 1968 pandemic strains are of avian origin. Unlike H2N2, the H3N2 subtype is still in circulation, however the high rate of antigenic drift of human H3 coupled with the long interval since the previous pandemic may mean that the human herd would have 'forgotten' the antigenic structure of the 1968 pandemic strain and therefore the reemergence of a similar strain circulating in the avian or swine reservoir could have potentially damaging consequences. Identifying such strains is of paramount value for pandemic surveillance and preparedness.

To address this question in this study we measure the 'antigenic intactness' of HA from avian or swine species in reference to HA from the corresponding pandemic subtypes. The antigenic identity (AI) of an avian or a swine HA is defined by the percentage fraction of amino acids in the immunodominant antigenic sites that are conserved in the corresponding pandemic HA (H1, H2 and H3 subtype). The AI value varies between 0 and 100. Values closer to 100 indicate a high antigenic identity with the pandemic HA.

Experimental Methods

Calculation of AI values for H1, H2 and H3 subtypes.

AI values were calculated using the characterized antigenic sites of H1, H2 and H3 HA. For H1, 128, 129, 158, 160, 162, 163, 165, 166, 167 (Sa); 156, 159, 192, 193, 196, 198 (Sb); 140, 143, 145, 169, 173, 182, 207, 224, 225, 240, 273 (Ca); 78, 79, 81, 82, 83, 122 (Cb) were used. For H2, 162, 248 (I-A); 137, 187 (I-B); 131, 222, 218 (I-C), 80, 200 (I-D); 40 (II-A), 273 (II-B) were used. For H3, 122, 133, 137, 143, 144, 145, 146

(A); 155, 186, 188, 189, 193 (B); 53, 54, 275, 278 (C); 201, 205, 207, 208, 217, 220 (D); 62,78,81,83 (E) were used. Positions are numbered according to H3 molecule. The antigenic identity (AI) of an avian or a swine HA is defined by the percentage fraction of amino acids in the dominant antigenic sites that are conserved in the corresponding pandemic HA for each of the H1 (A/South Carolina/1/18), H2 (A/Albany/6/58(H2N2)) and H3 (A/Aichi/2/1968 (H3N2)) subtypes.

In silico identification of glycosylation sites.

Glycosylation sites are defined by the motif N-X-T/S, where X is any amino acid except Proline. A position in a HA amino acid sequence is considered to be glycosylated if it contains the N-X-T/S motif and is predicted by GlyProt (http://www.glycosciences.de/modeling/glyprot/php/ main.php) – an online tool for in silico glycosylation of proteins.

Cloning, baculovirus synthesis, recombinant expression and purification of representative H3 HAs.

Soluble versions (lacking membrane proximal C-terminus region) of HA from representative H3N2 swine isolates A/swine/Chonburi/05CB2/ 2005 and A/swine/Nakhon pathom/NIAH586-2/2005 were recombinantly expressed (with C-terminal His-tag) as described previously[193]. These representative H3 HAs had high AI values and the prototypic Leu226 and Ser228 residues characteristic of human-adapted H3 HAs. Briefly, recombinant baculoviruses with the HA gene were used to infect (MOI51) suspension cultures of Sf9 cells (Invitrogen, Carlsbad, CA) cultured in BD Baculogold Max-XP SFM (BD Biosciences, San Jose, CA). The infection was monitored and the conditioned media was harvested 3–4 days post-infection. The soluble HA from the harvested conditioned media was purified using Nickel affinity chromatography (HisTrap HP columns, GE Healthcare, Piscataway, NJ). Eluting fractions containing HA were pooled, concentrated and buffer exchanged into 1X PBS pH 8.0 (Gibco) using 100K MWCO spin columns (Millipore, Billerica, MA). The purified protein was quantified using BCA method (Pierce).

Glycan array analysis.

To investigate the multivalent HA-glycan interactions a streptavidin plate array comprising of representative biotinylated $\alpha 2 \rightarrow 3$ and $\alpha 2 \rightarrow 6$ sialylated glycans was used as described previously [193]. 3'SLN, 3'SLN-LN, 3'SLN-LN-LN are representative avian receptors. 6'SLN and 6'SLN-LN are representative human receptors. The biotinylated glycans were obtained from the Consortium of Functional Glycomics through their resource request program. Streptavidin-coated High Binding Capacity 384-well plates (Pierce) were loaded to the full capacity of each well by incubating the well with 50 μl of 2.4 μM of biotinylated glycans overnight at 4 degrees C. Excess glycans were removed through extensive washing with PBS. The trimeric HA unit comprises of three HA monomers (and hence three RBS, one for each monomer). The spatial arrangement of the biotinylated glycans in the wells of the streptavidin plate array favors binding to only one of the three HA monomers in the trimeric HA unit. Therefore, in order to specifically enhance the multivalency in the HA-glycan interactions, the recombinant HA proteins were pre-complexed with the primary and secondary antibodies in the molar ratio of 4:2: 1 (HA: primary: secondary). The identical arrangement of 4 trimeric HA units in the pre-complex for all the HAs permit comparison between their glycan binding affinities. A stock solution containing appropriate amounts of Histidine tagged HA protein, primary antibody (Mouse anti 6X His tag IgG) and secondary antibody (HRP conjugated goat anti-Mouse IgG (Santacruz Biotechnology) in the ratio 4:2:1 and incubated on ice for 20 min. Appropriate amounts of pre-complexed stock HA were diluted to 250 μl with1% BSA in PBS. 50 μl of this pre-complexed HA was added to each of the glycan- coated wells and incubated at room temperature for 2 hours followed by the above wash steps. The binding signal was determined based on HRP activity using Amplex Red Peroxidase Assay (Invitrogen, CA) according to the manufacturer's instructions. The experiments were done in triplicate. Minimal binding signals were observed in the negative controls including binding of pre-complexed unit to wells without glycans and binding of the antibodies alone to the wells with glycans.

Results

We first applied the AI metric to human-adapted H1N1 and avian H2 subtypes for two reasons. In the former case, we tested the ability of AI values to discriminate the 1918 and 2009 pandemic HAs from the seasonal strains. In the latter case, we validated AI's potential to highlight the conservation of antigenic sites in avian H2[187]. For H1N1, the HA of the human-adapted strains were compared to 1918 pandemic H1N1 HA (A/South Carolina/1/18) and the characterized H1 antigenic sites Sa, Sb, Ca, Cb[194] [195] were used to calculate AI (Methods). The AI values clearly discriminate the reemerging swine-origin HA of 2009 H1N1 pandemic from the seasonal H1 based on the antigenic identity to the 1918 pandemic H1N1 HA (Figure

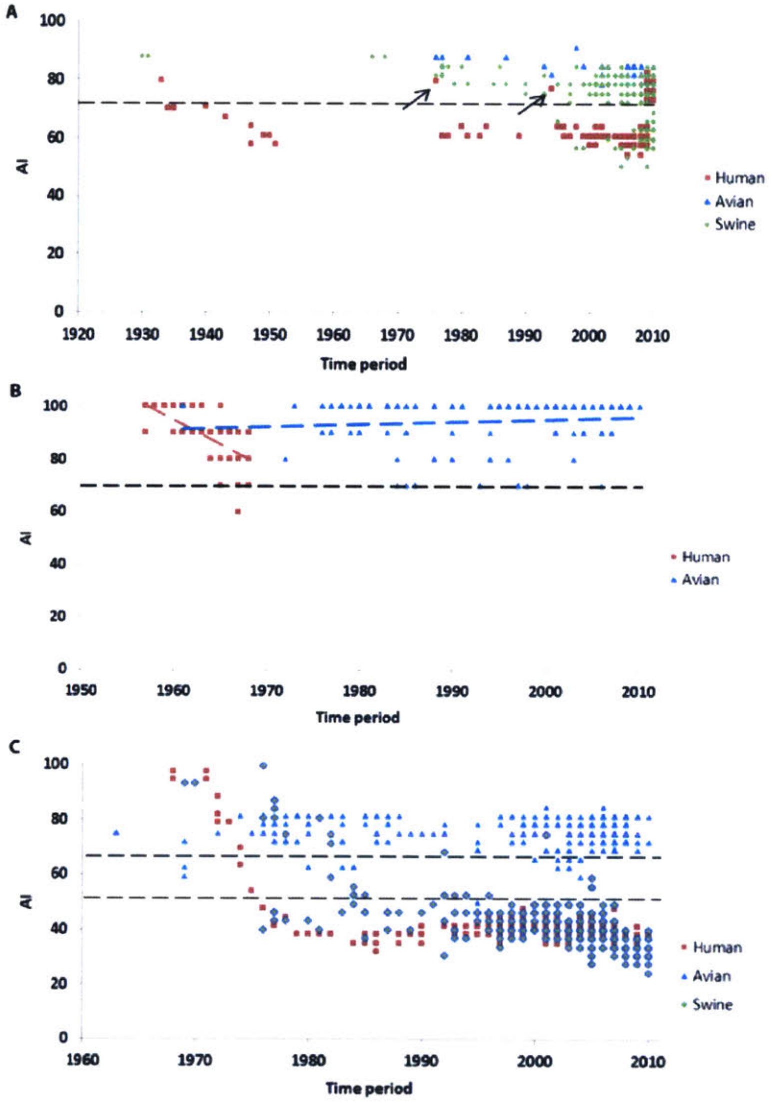


Figure 5.1 **Antigenic identity of HA from human, avian and swine species relative to pandemics that of note:** (a) 1918–19 H1N1, (b) 1957–58 H2N2, (c) 1968–69 H3N2 plotted against time of isolation (x-axis). To generate the plots, 2, 927 human, 166 avian, 950 swine HAs of H1 subtype; 117 human and 163 avian HAs of H2 subtype; 3,632 human, 756 avian and 347 swine HAs of H3 subtype were used. The two black arrows in (a). correspond to A/New Jersey/76 (H1N1) and A/Wisconsin/4754/1994 (H1N1), both of which caused human infections following pig-human interspecies transmission and have high AI values similar to the 2009 pandemic H1N1 strains. Dotted trendlines are added to graphically display the antigenic drift in avian vs. human H2 (b). The slope of the avian H2 trendline is 0.0845, whereas the slope of the human H2 trendline is −1.866. The dotted horizontal line indicates cutoff AI values (70% (H1); 70% (H2); 70% & 49% (H3)). The data points were jittered slightly on y-axis to avoid large overlaps (AI ~ AI + ε, where −1< ε <1).

5.1A). The reemerging swine-origin HA of the 2009 H1N1 pandemic and those that circulated during the 1918–40 period are characterized by AI values > 70% and markedly differ from the strains that circulated during 1940–2008 (varies from 48% to 77% with an average of 55%). Two 'classical' swine viruses (data points marked by black arrows), A/New Jersey/76 (H1N1) and A/Wisconsin/4754/1994 (H1N1), isolated between 1940–2008, also have high AI value and are genetically distinct when compared to the main cluster of human influenza viruses circulating in that period. Both viruses are known to have caused human infections following pig-human interspecies transmission. The A/New Jersey/76 influenza virus is reported to have caused respiratory illness in 13 soldiers with 1 death at Fort Dix, New Jersey[196]. The A/Wisconsin/4754/1994 virus was recovered from a 39 year-old man who came in close contact with experimentally infected pigs [197]. For the H2 subtype, the HA of the avian H2 strains were compared to the 1957–58 pandemic H2N2 HA (A/Albany/6/ 58(H2N2)) and the antigenic sites I-A, I-B, I-C, I-D, II-A and II- B [198] characterized by hybridoma antibodies generated in BALB/c mice were used to calculate AI. Consistent with the

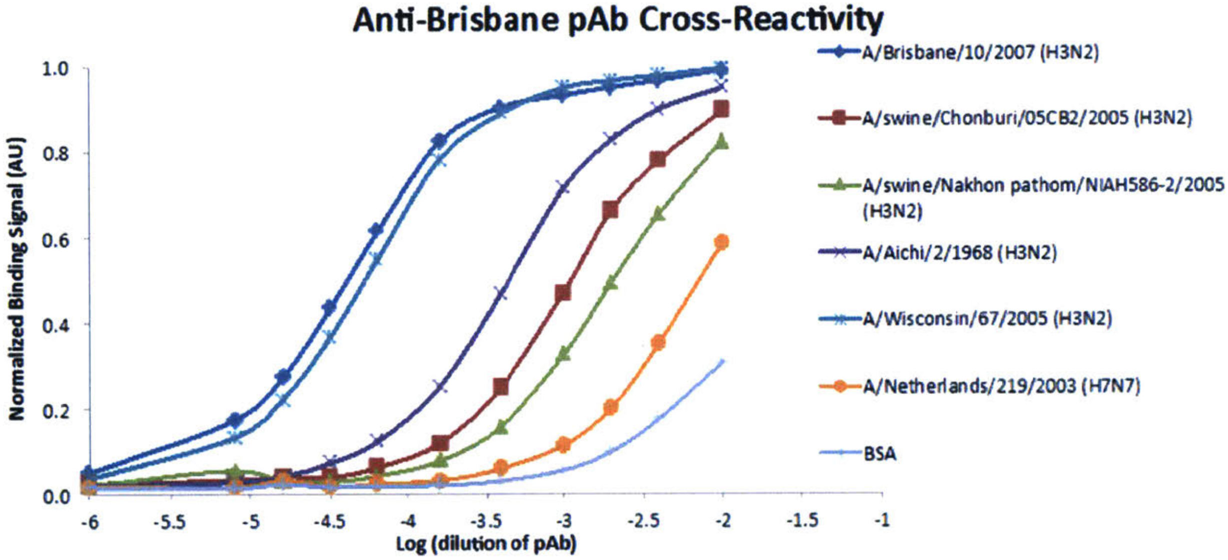


Figure 5.2 **Binding of anti-A/Brisbane/10/2007(H3N2) pAb to H3 strains measured by ELISA.** Tested were 2 swine H3N2 HAs (A/swine/Chonburi/05CB2/2005, A/swine/Nakhon pathom/NIAH586-2/2005), pandemic H3N2 HA (A/Aichi/2/1968), 1 seasonal H3N2 HA (A/Wisconsin/67/2005). The seasonal vaccine H3N2 HA (A/Brisbane/10/2007) and a representative H7N7 HA (A/Netherlands/219/2003) were used as positive and negative controls, respectively.

findings of a previous report [187], the AI values indicate that the antigenic sites of the 1957–58 pandemic H2N2 HA are conserved in circulating avian H2 influenza viruses (Figure 5.1b). In fact, the antigenic sites of the majority of avian H2 viruses in circulation are 100% identical to the 1957–58 pandemic H2N2 HA (Figure 5.1b). The conservation of antigenic sites in swine H2 influenza could not be assessed using this method due to lack of sequence information (H2N2 viruses do not circulate in swine; indeed, infection of swine with H2 viruses is rarely recorded). Similar to H1 subtype, the evolution of human H2 is characterized by steady antigenic drift leaning away from the pandemic strain. Although the majority of the viral strains that circulated during the immediate post-pandemic period 1957–68 have AI values >70%, viral strains with AI~60% appeared after 1967 (Figure 5.1b). It is reasonable to expect that the AI values would have decreased further had H2 continued to circulate in human population as a seasonal virus after 1968. The above analyses using H1 and H2 subtypes suggest that viruses carrying pandemic HA-like genes can be distinguished from seasonal viruses using a cutoff value AI= ~70%.

	CA/09	SD/03	Perth/09	KS/09	PA/10	WI/10	MN/10
A/California/07/2009 (H1N1pdm09)		70.71	0.78	0.39	1.1	0.55	0.55
A/South Dakota/03/2008 (Human H1N1-SOIV)	81.48		0.55	1.1	1.56	0.39	0.39
A/Perth/16/2009 (Seasonal H3N2)	22.22	22.22		0.78	50	1.1	1.1
A/Kansas/13/2009 (Human H3N2-SOIV)	22.22	22.22	66.67		17.68	6.25	3.13
A/Pennsylvania/14/2010 (Human H3N2-SOIV)	18.52	18.52	66.67	88.89		35.36	35.36
A/Wisconsin/12/2010 (Human H3N2-SOIV)	22.22	22.22	66.67	88.89	92.59		35.36
A/Minnesota/11/2010 (Human H3N2-SOIV)	22.22	22.22	66.67	88.89	92.59	100	

Table 5.1 **HI-based antigenic relatedness** (upper right) **and AI values** (lower left) in pairwise comparisons among 7 influenza H3N2 viruses isolated from 2008 to 2010 (R = 0.603314, p-value = 0.002)

In the case of H3, the 5 antigenic sites (A–E)[199] [59] were used to calculate AI in reference to the prototypic pandemic strain of 1968 (A/Aichi/2/1968 (H3N2)). Unless stated otherwise henceforth an AI value for a given H3 HA sequence refers to its antigenic identity with the 1968 pandemic H3 HA. A total of 1,103 H3 avian and swine sequences were downloaded from the NCBI Influenza Database and analyzed. Of these 1,103 sequences, 756 were of avian origin and 347 were of swine origin. The avian sequences comprised nine different subtypes (H3N1-9), and the swine sequences comprised four different subtypes (H3N1, H3N2, H3N3 and H3N8). The avian and swine H3 amino acid sequences were compared against A/Aichi/ 2/1968 and AI values were computed for all the sequences (Figure 5.1c). In addition, a total of 3,632 human-adapted H3N2 HA sequences were downloaded from the NCBI database and compared against 1968 pandemic H3 HA to enable a cross-species comparison of the antigenic drift (Figure 5.1c). The AI values and phylogeny analysis indicate that, in comparison with recent human H3, avian and swine H3 are genetically and antigenically closer to the 1968 pandemic HA. Thus, we confirmed that avian and swine H3 are indeed antigenically intact (Figure 5.1c).

In addition to the amino acids that constitute the antigenic sites, the attachment of complex glycans at specific glycosylation sites (Asn-X-Ser/Asn-X-Thr, where X is not a Proline) is also often part of the antigenic surface. An increase or decrease in the number of N-glycosylation sites therefore critically governs the antigenic properties of HA. The 1968 pandemic H3 HA carries only two glycosylation sites on the globular head region (at 81 & 165), whereas HA from seasonal strains carries an average of six sites (at 63, 122, 126, 133, 144, 165) [200]. To incorporate glycosylation in the calculation of antigenic identity, the globular head region of the avian and swine HA sequences were examined for the conservation of 1968 pandemic H3-like glycosylation pattern (experimental methods). Among the 1,103 avian and swine H3 HA sequences, 359 carried additional glycosylation sites or positional shifts and therefore were removed from further consideration. The remaining 744 HA sequences (~ 67%) were found to possess the 1968 pandemic HA-like glycosylation pattern. Out of the 744 HA sequences, strains corresponding to 449 sequences (all avian) were isolated after 2000—many as recently as 2010—and their AI value exceed 70%.

	HK/71	ENG/72	PC/73	MC/75	VIC/75	TOK/75	ENG/75	BAN/1/79	BAN/2/79
A/Hong Kong/107/71		3.61	5.1	2.55	1.81	2.55	2.08	1.47	0.9
A/England/42/72	66.67		25	3.83	6.25	1.56	0.64	1.28	0.55
A/Port Chalmers/1/73	74.07	77.78		12.5	6.25	3.13	3.61	1.81	1.1
A/Mayo Clinic/1/75	59.26	62.96	81.48		10.87	3.13	2.21	1.81	1.1
A/Victoria/3/75	51.85	55.56	77.78	81.48		8.85	3.61	1.28	0.78
A/Tokyo/1/75	74.07	62.96	74.07	70.37	66.67		1.28	2.55	1.1
A/England/864/75	51.85	55.56	77.78	85.19	88.89	62.96		14.49	5.1
A/Bangkok/1/79	33.33	48.15	59.26	59.26	62.96	37.04	74.07		10.87
A/Bangkok/2/79	33.33	51.85	55.56	55.56	51.85	37.04	62.96	88.89	

Table 5.2 **HI-based antigenic relatedness** (upper right) **and AI values** (lower left) in pairwise comparisons among 9 influenza H3N2 viruses isolated from 1970 to 1979 (R = 0.523472, p-value = 0.00057).

Extrapolating from H1 and H2 pandemic scenarios, the above strains are likely to pose a threat should they acquire the mutations necessary to crossover into human population. Of note, a novel H3N8 avian influenza virus acquired the ability to infect harbor seals in New England recently [201]. The AI of the seal H3N8 HA is 78%, which is the habitual AI range of avian H3 influenza viruses. Given the high AI value, the history of the spread of avian influenza to humans and the fact that seal H3N8 has already acquired potential to bind sialic acid receptors that are commonly found in the mammalian respiratory tract [201], seal H3N8 virus could jump, directly or via reassortment, to humans with pandemic consequences. More recently, the CDC reported the outbreak of a triple reassortant H3N2 swine-origin influenza virus (SOIV) and released a set of sequences at Global Initiative on Sharing All Influenza Data (GISAID) following this event. The HA of a prototype outbreak strain, A/Minnesota/11/2010 (referred as Minn10), shares very high homology (approx. 98%) with the HA of swine A/swine/Minnesota/ 7931/2007(H3N2) (SwMinn10), and has good binding and transmission properties [202].

Although the AI value of SwMinn10 (approx. 39%) is comparable to that of a typical seasonal H3 HA, they share very low antigenic identity between them (only 15 out of the 27 [approx. 55%] antigenic positions are conserved). More importantly, the glycosylation pattern appears to be very different between SwMinn10 and seasonal H3 HA. The SwMinn10 HA contains only three glycosylation sites in the globular head region, compared to 6 for the seasonal HA. The swine predecessor was not part of the 581 sequences identified by the analysis. This is due to its low AI value and the extra (third) glycosylation site in the head region. Although Minn10 cannot be regarded as a strain resembling the 1968 pandemic strain, the outbreak caused by this virus supports our theory that avian and swine strains that are divergent enough from the seasonal HA, both antigenically and with respect to their glycosylation pattern, need to be considered as potential threats. Consequently, based on the above observations, we relaxed the criteria used to identify potential pandemic strains and considered those HAs isolated after 2000, having matching glycosylation pattern as pandemic H3 and whose AI was equal to or greater than 49%, the maximum AI value of recent seasonal H3 (2000 or after) (Figure 5.1c). This yielded 581 sequences (549 avian, 32 swine). If a virus carrying a HA similar to any one of the 581 sequences acquires the potential to crossover into humans, it would likely have a major impact on both immune recognition and vaccine efficacy. The efficacy of the influenza vaccine in humans is thought to correlate well with the 'antigenic relatedness' metric (reciprocal of antigenic distance) obtained from ferret antisera hemagglutinin inhibition (HI) assays [203] [204] between the vaccine strain and the circulating epidemic strains[205] [206]. We tested the degree of correlation between the AI values and the HI-derived antigenic relatedness to: (1) assess the potential of AI in predicting vaccine-induced cross-reactive antibody responses; and (2) to evaluate the cross-protective capacity of the current vaccine strain, A/Victoria/361/2011 (H3N2), against potential threats. For this exercise, we analyzed three sets of ferret serum HI cross-reactivity data where amino acid sequences of the HA1 polypeptide were present.

	JHB/94	Wuh/95	NC/95	Syd/97	Mosc/99	Pan/99
A/Johannesburg/33/94		8.84	8.84	2.21	3.13	3.13
A/Wuhan/359/95	92.59		50	3.13	3.13	4.42
A/Nanchang/933/95	92.59	100		3.13	70.71	4.42
A/Sydney/5/97	81.48	88.89	88.89		70.71	70.71
A/Moscow/10/99	81.48	81.48	91.48	92.59		100
A/Panama/2007/99	77.78	85.19	85.19	92.59	92.59	

Table 5.3 **HI-based antigenic relatedness** (upper right) **and AI values** (lower left) in pairwise comparisons among 6 influenza H3N2 viruses isolated from 1994 to 1999 (R = 0.61, p-value = 0.007)

The first set contained 7 viral strains (21 pair- wise comparisons) isolated from 2008 to 2010. The second set contained 9 viral strains (36 pairwise comparisons) isolated from 1970 to 1979[207]. The third set contained 6 viral strains (15 pairwise comparisons) isolated from 1994 to 1999[208]. Antigenic relatedness between two viral strains based on ferret anti-serum was determined using the method described by Lee[204]. Briefly, the antigenic relatedness between two viral strains is directly proportional to the ratio of the product of the heterologous titers against each other to the product of the homologous titers. In total, 72 pairwise comparisons among 22 viruses were available for analysis. Among the 72 pairwise comparisons, 5 (7%) have an antigenic relatedness >70% (i.e., similar antigenicity), and 67 (93%) have an antigenic relatedness <70% (i.e., antigenic variant).

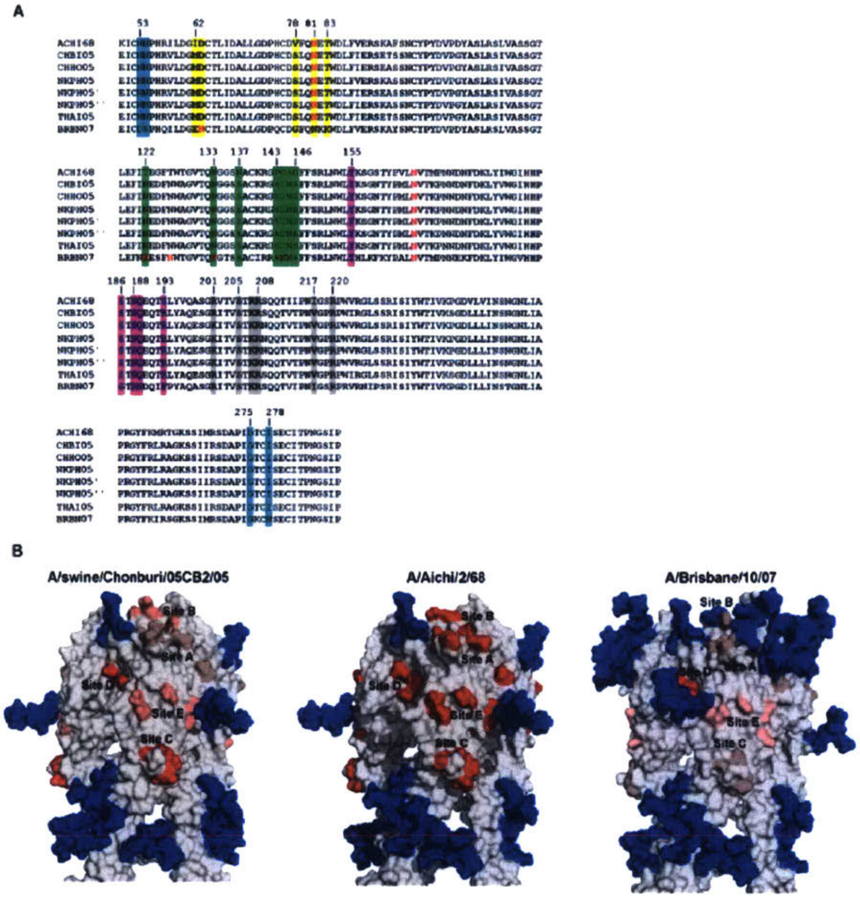


Figure 5.3 **Genetic, antigenic and glycosylation-pattern relatedness of 1968 pandemic H3N2 HA to seasonal, swine and avian H3 HA.**(a) Sequence alignment of the expanded globular head region (residues 50–328) of the HAs listed in Figure 5.2. Antigenic sites A, B, C, D, E of H3 HA are highlighted in green, magenta, cyan, grey and yellow, respectively. In each sequence, the Asn residue associated with the N-linked glycosylation sites (Asn-X-Ser/Asn-X/Thr) is marked in red. (b) Surface rendered three-dimensional structural models of trimeric HA1 globular head of representative pandemic (middle), seasonal (right) and swine (left) HAs. The view of trimer is along axis perpendicular to 3-fold symmetry axis to give a complete picture of the antigenic and glycosylation sites. The antigenic sites A–E are marked on the structure. With 1968 pandemic HA as reference (antigenic sites shown in red), the structural similarity of the antigenic sites in seasonal and swine HAs to the reference HA is shown in different shades of red (duller shade representing low similarity to brighter shade representing high similarity).

Results indicate that the AI values have significant correlation with the HI-based antigenic relatedness metric (Table 5.1 , Table 5.2, Table 5.3), indicating that AI values could be applied to predict vaccine-induced cross-reactive antibody responses and thus selection of vaccine strains. Particularly, antigenically related viral strains (>70%) have AI>80%, hence we employed an 80% cutoff to determine protection, or lack thereof, between a vaccine strain and a challenge viral strain. The current H3N2 vaccine strain A/Victoria/ 361/2011 has AI values of 92% with the seasonal viral strain A/Brisbane/10/2007, 29% with the A/Aichi/2/68, 56% with a typical H3N2 SOIV and 44% with a representative swine H3 strain from the group of swine viral strains having AI>49%. These data indicate that the current vaccine strain is unlikely to offer cross-protection against the circulating swine or SOIV viruses whatsoever. Supporting this, IgG polyclonal raised in rabbit with seasonal vaccine H3 strain (A/Brisbane/10/2007(H3N2)) preferentially bind to current seasonal H3 but have weaker affinity to a representative swine H3 (Figure 5.2). More significantly, out of the 581 HA sequences, six swine HAs already contain the prototypic mutations (L226, S228) necessary for HA human adaptation [209], and are thus capable of entering the human population either directly or via reassortment (Table 5.4, Figure 5.3) [209]. We recombinantly expressed HA derived from two swine isolates, A/swine/Chonburi/05CB2/2005 (H3N2) and A/swine/ Nakhon pathom/NIAH586-2/2005 (H3N2), which have high AI value (Table 5.4) and characterized their relative binding affinities to representative avian and human receptors on a glycan array platform (Methods, Figure 5.4). Both swine HAs showed high affinity binding to both human and avian receptors. The high affinity human receptor- binding of these swine HAs appears to be in the same range as that of other seasonal H3 HAs characterized previously [15] [193], and are thus capable of entering the human population either directly or via reassortment. The antigenic relationship of these HAs (AI value and glycosylation pattern) to the pandemic 1968 H3N2 HA strongly suggests that the six isolates belong to swine virus lineage and not examples of transient reverse zoonoses. Phylogenetic analysis of the 32 swine isolates revealed that majority of them fall under European and Asian swine lineages.

The analyses presented here portend a vaccine strategy to prevent a future H3 pandemic. Among the WHO recommended vaccine strains of influenza A/H3N2 virus, A/Hong Kong/1/1968 (H3N2) will be effective (AI > 80%) against 505 of 581 strains (~ 87%) identified by this study, and thus could be used for the development of pandemic influenza vaccine. Surprisingly, H3N2 vaccine strains that were subsequently used are not capable of being as effective. These data suggest that a cocktail of A/Hong Kong/1/1968 (H3N2) and an avian and swine strain each that represent the circulating influenza in birds and pigs can form the components of the pandemic influenza vaccine. To understand the results from AI calculations in the context of the spatial relationship between glycosylation site and antigenic sites of H3 HA we constructed structural homology models of HA1 globular head of ACHI68, BRBN07 and CHIB05 HAs (see Table 5.4 for strain information). These structural models of HA comprised the basic trimannosyl core of N-linked glycan attached to the glycosylation sites (Figure 5.3). From the structural comparison it is clear that antigenic shape of HA which includes antigenic sites A-E and the glycosylation pattern of HA1 from the swine strain (CHIB05) closely resembles that of the 1968

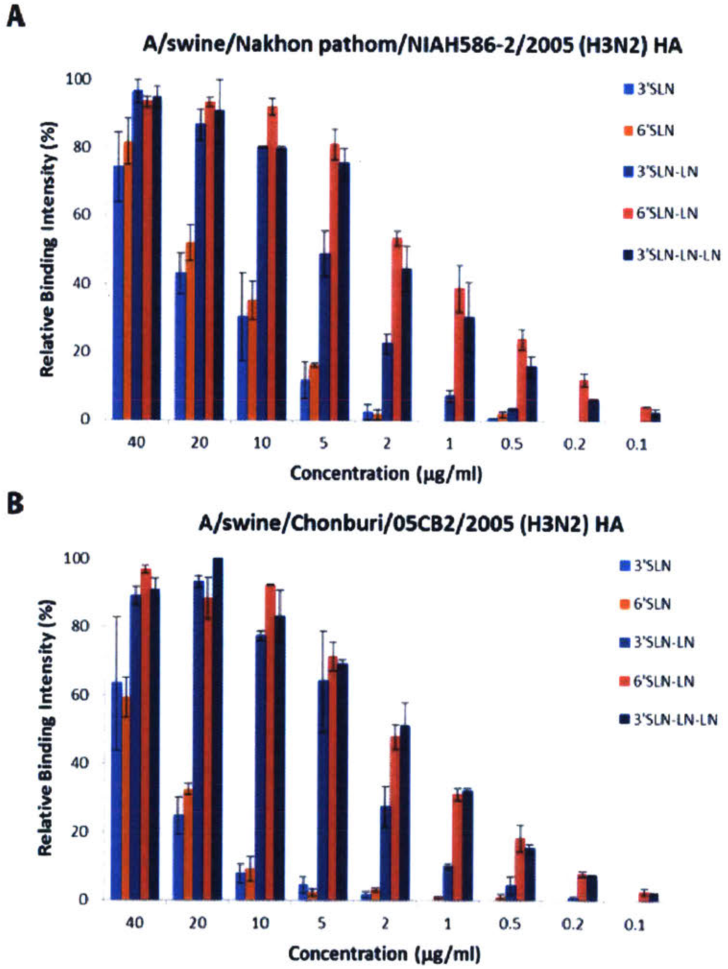


Figure 5.4 **Glycan microarray analysis of representative H3 HAs.** Dose dependent binding of A/swine/Nakhon pathom/NIAH586-2/2005 (a), and A/swine/Chonburi/05CB2/2005 (b) HAs to representative avian and human glycan receptors on the glycan array platform is shown. Both these HAs show high affinity binding to both human receptors (6′SLN-LN) and avian receptors (3′SLN-LN and 3′SLN-LN-LN).

pandemic HA. Conversely, the antigenic shape of a more recent seasonal strain (BRBN07) is remarkably different from that of the pandemic strain.

ACCESSION	VIRUS NAME	ABBREVIATION	AI%	GLYCOSYLATED POSITIONS
AAA43239	A/Aichi/2/1968(H3N2)	ACHI68	100	81, 165
ABY40417	A/swine/Chonburi/05CB2/2005(H3N2)	CHBI05	56	81, 165
ABY40412	A/swine/Chachoengsao/NIAH586/2005(H3N2)	CHHO05	52	81, 165
ABY40414	A/swine/Nakhon pathom/NIAH586-2/2005(H3N2)	NKPH05	52	81, 165
BAH02120	A/swine/Nakhon pathom/NIAH586-1/2005(H3N2)	NKPH05'	52	81, 165
ABY40413	A/swine/Nakhon pathom/NIAH586-1/2005(H3N2)	NKPH05"	52	81, 165
ACM80372	A/swine/Thailand/S1/2005(H3N2)	THAI05	56	81, 165
ABW23353	A/Brisbane/10/2007(H3N2)	BRBN07	37	63, 122, 126, 133, 144, 165
ACS71642	A/Perth/16/2009(H3N2)	PTH09	37	63, 122, 126, 133, 165

Table 5.4 **Avian and Swine HAs antigenically similar to 1968 pandemic H3N2 HA**. The AI values and glycosylation pattern of A/Aichi/2/1968(H3N2) and six swine HAs having prototypic mutations (L226, S228) necessary for HA human adaptation are compared alongside. A representative seasonal vaccine strain (A/Brisbane/10/2007(H3N2) and A/Perth/16/2009(H3N2)) are included to show the variation in the AI values and glycosylation pattern

Conclusion

The H3 HA of some of the recent avian strains share approximately 86% overall sequence identity with the HA of the avian progenitor of the 1968 pandemic virus (A/duck/Ukraine/1/1963), reflecting antigenic intactness within birds. Many sequences from swine, some collected as recently as 2001, were also found to have high homology with the A/duck/Ukraine/1/1963HA, indicating avian to swine transfer. For reasons that remain unclear, the more recent swine H3 HAs (2006 and later) have diverged significantly from the 1968 pandemic H3N2 HA (Figure 5.1c) while in contrast the majority of swine H1 HAs remained antigenically stable from1918 to the 1990s. Unlike the 1918 H1N1 virus which crossed to swine soon after and remained in swine, the human H3N2 viruses have repeatedly crossed from humans to swine for some time – quite possibly, this could be the reason why swine H3 viruses appear to manifest the antigenic drift that human strains underwent during this period. In fact, the AI values of human H3 in the last decade are comparable to the AI values of some swine H3 HAs of the same period; interestingly, the recent human and swine HAs show differential binding to polyclonal antibodies generated against seasonal vaccine strain (Figure 5.2). This apparent discrepancy may be explained in part by the presence of certain key antigenic "hotspot" locations, where amino acid substitutions can lead to disproportionately large changes in antigenicity. Our observation is supported by other studies on H3 antigenic evolution[54] [210]. The frequent interspecies transmission of H3 viruses might also explain why this subtype is associated with the highest rates of mortality[211].

The importance of glycosylation in antigenic site masking leading to a new pandemic cycle and viral evolution became apparent after the 2009 pandemic. It was observed that the seasonal H1N1 HA carries antigenic site-masking glycosylation sites not present in the 2009 pandemic H1N1 HA (and 1918 H1N1 HA) and the exposure of the unprotected antigenic surface is believed to be the reason underpinning the severity of the 2009 H1N1 pandemic. Akin to H1 subtype, the additional glycosylation sites on the recent seasonal H3 appear to have a role in antigenic site-masking. For instance, the glycosylation at position 63 masks antigenic site E, and glycosylation at sites 122, 133, and 144 protect antigenic A. The shielding nature of these glycosylation sites is evident from the gradual decline in the mutation rate of the masked antigenic sites following their appearance, portending a 2009 H1N1-like H3N2 pandemic. If a virus carrying a HA similar to the ones identified by this analysis makes its way into humans, it would need to evolve rapidly in response to selective pressures from vaccination and herd immunity. The ability of H3 subtype to add glycosylation sites will be a key factor enabling the virus to achieve sustained circulation in the next cycle. In contrast, a previous study [212] using nucleotide sequence analysis concluded that H2 has an intrinsically lower capacity to add glycosylation sites. Taking these factors together, we assert that it is less likely for an avian or swine H2 virus (antigenically similar to 1957–58 pandemic H2N2) to gain a foothold for sustained circulation in humans when compared to H3 viruses. The rapid antigenic drift that human H3N2 HA underwent during the early adaptation period of the virus (1968–76) appears to have slowed down after 1977 (Figure 5.1c). Interestingly, this time period also coincides with the reemergence of H1N1 in the human population. The (re-) emerging H1N1 subtype could have imposed strong selective pressures on the H3N2 to stop circulating in humans after 1977. The evolution of human H3N2 HA after 1977 is characterized by glycosylation accrual, low-level site-specific antigenic changes, and variations at other non-immunodominant sites (Figure 5.1 c). Additionally, a recent study found that the affinity of human H3 viruses for human receptors has reduced drastically since 2001[144]. These observations suggest that currently circulating viruses are not as dominant as the earlier viruses. Based on this trend, one can argue that human H3N2 HA presently is "antigenically drained", which poses a substantially high barrier to evolution via antigenic drift. However, the presence of antigenically intact H3 in avian and swine suggests that, as with 2009 H1N1 pandemic, reassortment can result in 'resetting and shifting' the antigenicity back to that of the 1968 pandemic and hence facilitate sustained evolution of this subtype in humans. Influenza A viruses of other subtypes (H5, H7, H9) that have caused sporadic infections in humans over the past decade also pose equal risk of a pandemic, especially since they represent completely novel HA subtypes. Although antigenic phenotypes could be predicted from HA sequences, the genetic signatures in influenza viruses that lead to a sustained human-to-human transmission cannot be accurately predicted. Although an antigenically novel HA is necessary, it is not the only determining factor for a pandemic. While gain of host receptor specificity is a key determinant, changes in influenza proteins other than HA such as the polymerase (PB2) are typically involved, making predictions of the timing of future pandemics more complex. Nevertheless, our study facilitates setting the stage for future work aimed at designing vaccination studies with animal models using a cocktail of H3 antigens from strains of current avian and swine origin along with specific past strains. Such studies would augment the preparedness in the event of potential re-emergence of H3N2 pandemic [183].

The 2009 swine-origin H1N1 influenza, though antigenically novel to the population at the time, was antigenically similar to the 1918 H1N1 pandemic influenza, and consequently was considered to be "archived" in the swine species before reemerging in humans. Given that the H3N2 is another subtype that currently circulates in the human population and is high on WHO pandemic preparedness list, we assessed the likelihood of reemergence of H3N2 from a non-human host. Using HA sequence features relevant to immune recognition, receptor binding and transmission we have identified several recent H3 strains in avian and swine that present hallmarks of a reemerging virus. IgG polyclonal raised in rabbit with recent seasonal vaccine H3 fail to recognize these swine H3 strains suggesting that existing vaccines may not be effective in protecting against these strains. Vaccine strategies can mitigate risks associated with a potential H3N2 pandemic in humans.

This work resulted in the following publication:
"Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic" published May 10, 2013 in Scientific Reports,
Article number: 1822; DOI: doi:10.1038/srep01822

Abstract

The 2009 swine-origin H1N1 influenza, though antigenically novel to the population at the time, was antigenically similar to the 1918 H1N1 pandemic influenza, and consequently was considered to be "archived" in the swine species before reemerging in humans. Given that the H3N2 is another subtype that currently circulates in the human population and is high on WHO pandemic preparedness list, we assessed the likelihood of reemergence of H3N2 from a non-human host. Using HA sequence features relevant to immune recognition, receptor binding and transmission we have identified several recent H3 strains in avian and swine that present hallmarks of a reemerging virus. IgG polyclonal raised in rabbit with recent seasonal vaccine H3 fail to recognize these swine H3 strains suggesting that existing vaccines may not be effective in protecting against these strains. Vaccine strategies can mitigate risks associated with a potential H3N2 pandemic in humans.

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Kannan Tharakaraman, Rahul Raman, Karthik Viswanathan, V. Sasisekharan and Ram Sasisekharan for their contributions to this work. This work was funded in part by the National Institutes of Health (R37 GM057073-13) and the National-Research-Foundation-supported Interdisciplinary Research group in Infectious Diseases of SMART (Singapore-MIT Alliance for Research and Technology).

Chapter 6 : New England harbor seal H3N8 influenza virus retains avian-like receptor specificity

Summary and Significance

In late 2011, roughly 200 New England harbor seals died in a pneumonia outbreak associated with an H3N8 virus. Sequence analysis of the H3N8 virus revealed that it possessed mutations in the polymerase gene (PB2) that were associated with human adaptation [213]. Furthermore, in chapter 5, a pandemic risk assessment framework was developed that measured the antigenic intactness of circulating H3s (i.e. measuring the degree of antigenic similarity between circulating H3s and the 1968 pandemic HAs). Analysis of the 2011 seal H3N8 revealed that it shared a high degree of similarity to the 1968 pandemic H3N2 HA [176]. This data suggested that H3N8, should it infect humans, would be antigenically novel and our recent vaccines would not provide sufficient protections. Finally, from an ecological perspective the increased urbanization of coastal cities, has given rise to increases in seal-human contact. Taken together, these features suggest that this seal H3N8 might be able to infect humans and thus poses a significant pandemic risk.

The goal of this chapter is to better understand the glycan binding specificity of the 2011 seal H3N8. Earlier studies investigating the glycan receptor binding properties of H3N8 HA determined that it bound both $\alpha 2\rightarrow 6$ and $\alpha 2\rightarrow 3$ linked glycan receptors [213,214]. However, this was measured using a hemagglutination assay and a solid phase binding assay, and did not account for critical features of the HA glycan interaction (i.e. multivalent presentation of the HA, or assessing HA binding to glycans adopting cone- and umbrella-like topology [15,41]). Towards this, 2011 Seal H3N8 HA was expressed and subjected to a glycan array to exhaustively characterize its interaction with over 600 unique glycan structures. Additionally, the binding to physiologic glycan receptors and its replication in human lung cells were measured.

Our results indicate that the seal H3N8 HA preferentially recognizes $\alpha 2\rightarrow 3$-linked glycans. Furthermore, H3N8 can replicate in human lung cancer cells. Thus, at present the seal H3N8 HA is, likely, unable to infect humans. However, given that it has the ability to replicate in human cells, if H3N8 HA acquired the high affinity binding to human glycan receptors it has potential to infect human lung tissue.

This work resulted in the following publication:

"New England harbor seal H3N8 influenza virus retains avian-like receptor specificity" published in *Scientific Reports* **6**, Article number: 21428 February 18 2016 ;
DOI: doi:10.1038/srep21428

Abstract

An influenza H3N8 virus, carrying mammalian adaptation mutations, was isolated from New England harbor seals in 2011. We sought to assess the risk of its human transmissibility using two complementary approaches. First, we tested the binding of recombinant hemagglutinin (HA) proteins of seal H3N8 and human-adapted H3N2 viruses to respiratory tissues of humans and ferrets. For human tissues, we observed strong tendency of the seal H3 to bind to lung alveoli, which was in direct contrast to the human-adapted H3 that bound mainly to the trachea. This staining pattern was also consistent in ferrets, the primary animal model for human influenza pathogenesis. Second, we compared the binding of the recombinant HAs to a library of 610 glycans. In contrast to the human H3, which bound almost exclusively to α-2,6 sialylated glycans, the seal H3 bound preferentially to α-2,3 sialylated glycans. Additionally, the seal H3N8 virus replicated in human lung carcinoma cells. Our data suggest that the seal H3N8 virus has retained its avian-like receptor binding specificity, but could potentially establish infection in human lungs.

This chapter is comprised of the following manuscript:

SCIENTIFIC REPORTS

OPEN

New England harbor seal H3N8 influenza virus retains avian-like receptor specificity

Received: 30 May 2015
Accepted: 25 January 2016
Published: 18 February 2016

Islam T. M. Hussein[1], Florian Krammer[3], Eric Ma[1], Michael Estrin[1], Karthik Viswanathan[2], Nathan W. Stebbins[1,2], Devin S. Quinlan[1,2], Ram Sasisekharan[1,2] & Jonathan Runstadler[1]

An influenza H3N8 virus, carrying mammalian adaptation mutations, was isolated from New England harbor seals in 2011. We sought to assess the risk of its human transmissibility using two complementary approaches. First, we tested the binding of recombinant hemagglutinin (HA) proteins of seal H3N8 and human-adapted H3N2 viruses to respiratory tissues of humans and ferrets. For human tissues, we observed strong tendency of the seal H3 to bind to lung alveoli, which was in direct contrast to the human-adapted H3 that bound mainly to the trachea. This staining pattern was also consistent in ferrets, the primary animal model for human influenza pathogenesis. Second, we compared the binding of the recombinant HAs to a library of 610 glycans. In contrast to the human H3, which bound almost exclusively to α-2,6 sialylated glycans, the seal H3 bound preferentially to α-2,3 sialylated glycans. Additionally, the seal H3N8 virus replicated in human lung carcinoma cells. Our data suggest that the seal H3N8 virus has retained its avian-like receptor binding specificity, but could potentially establish infection in human lungs.

Influenza A viruses (IAVs) caused several pandemics in the past and continue to pose significant threats to human public health[1]. Wild migratory birds are the natural reservoirs of IAVs from which IAV's occasionally cross the species barrier to infect domestic birds, humans and several other mammalian species[2]. Marine mammals are particularly interesting hosts for IAVs. They are globally distributed and can migrate over long distances in the vicinity of coastal ecosystems and population centers, where they intersect with waterfowl and shorebirds in scenarios conducive to virus exchange[3]. Cases of IAV infection in marine mammals have been documented in the literature with several IAV subtypes including H1N1, H3N3, H3N8, H4N5, H4N6, H7N7 and H10N7[4–8]. The majority of these transmission events have implicated an avian source; however, serological evidence for seal infection by human H3 viruses has also been reported[9–12]. A spillover of an H7N7 seal virus to humans has been also described[4,13]. The wide variety of IAV strains infecting seals provides opportunities for genetic reassortment and/or adaptation, and it has been proposed that seals might play a similar role to pigs as mixing vessels for avian and human viruses[14]. With the exponential increase in protected seal populations and urbanization of coastal cities, the seal-human interface is continuously expanding, which creates a suitable environment for viral zoonotic transmissions[15,16]. The recently isolated H3N8 (A/harbour seal/New Hampshire/179629/2011) virus from an outbreak in harbor seals (*Phoca vitulina*) in New England demonstrated naturally acquired polymerase mammalian adaptation mutations[17], indicating that it is of interest for human public health.

The viral surface glycoprotein, hemagglutinin (HA), is a key player in mediating transmission of IAVs. HA recognizes glycans with terminal sialic acid (SA) residues linked to galactose (Gal) via either an α-2,3 or α-2,6 linkage[18]. Glycan receptor binding specificity of IAVs helps define their host range and tissue tropism[19]. It is widely accepted that avian viruses preferentially bind to SAα-2,3Gal, while human influenzas bind to SAα-2,6Gal receptors[20,21]. Mutations switching HA's binding specificity to the α-2,6 SA linkage is likely an important step in establishing human transmissibility[22,23], with the overall glycan topology playing a critical role in determining receptor binding and host tropism[24].

As a step towards assessing the public health risks, this study provides a comprehensive assessment of the receptor-binding specificity of a recombinant seal H3N8 HA to physiological glycans displayed on human and

[1]Department of Biological Engineering and Division of Comparative Medicine, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Koch Institute of Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. Correspondence and requests for materials should be addressed to J.R. (email: jrun@mit.edu)

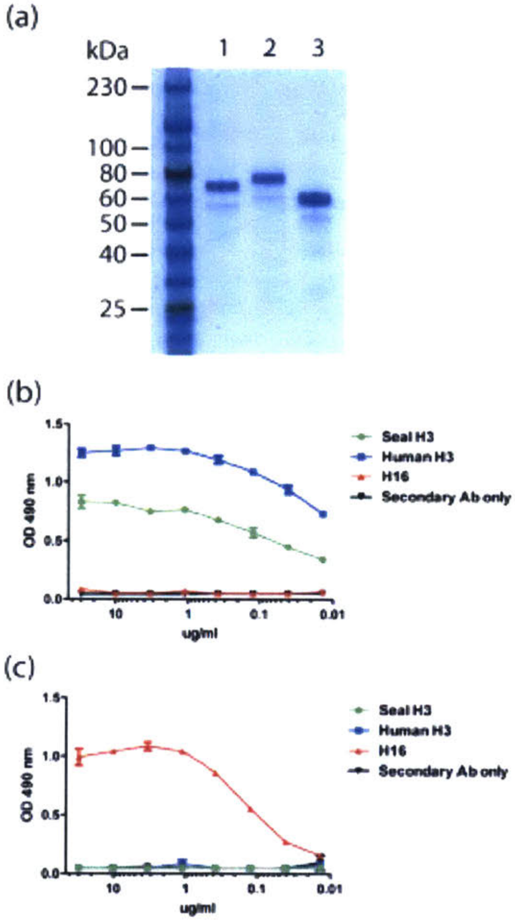


Figure 1. Expression and purification of recombinant HA proteins. (**a**) Coomassie blue stained SDS gel showing purified seal H3N8 HA (lane 1), human H3N2 HA (lane 2) and H16N3 (A/black headed gull/Sweden/5/1999) HA control (lane 3). (**b**) ELISA optical density values showing reactivity of the H3-specific monoclonal antibody 12D1 to recombinant seal H3 (green) and human H3 (blue) HA proteins. No reactivity was detected to the H16 control HA (red). (**c**) A control ELISA was performed using CR6261 antibody specific for group 1 HAs including H16, but not H3 subtype (which belongs to group 2). As expected CR6261 reacted with H16 HA (red), but showed no reactivity to the two H3 HAs.

ferret respiratory tissues, and to chemically synthesized glycan arrays. Seasonal human-adapted strains of influenza are known to bind to SAα-2,6Gal receptors[25], and are thus an epidemiologically relevant control for our study. In contrast to the human-adapted H3 control (A/Wyoming/03/2003), our findings suggest that the seal H3N8 HA preferentially binds to SAα-2,3Gal receptors that are abundant on human lungs[26]. We also present evidence that seal H3N8 virus replicated in human lung carcinoma cells, highlighting the importance of continuous monitoring of influenza viruses circulating in seals for the early detection of strains with enhanced zoonotic potential.

Results

Recombinant HA protein expression and purification. Large amounts of soluble trimeric HA proteins were produced through expression in insect cells using the baculovirus system. This system has been successfully used before to produce biologically active recombinant HA for structural and biochemical studies[27–28]. The purity and identity of expressed proteins were assessed by SDS-PAGE and ELISA. As shown in Fig. 1(a), Coomassie blue stained gel shows purified recombinant seal H3, human H3 and H16 control HAs. To confirm the identity of the recombinant HAs, each was tested by direct coating ELISA, where an H3-specific antibody (12D1) was found to react with both H3 proteins, but not an unrelated H16 control (Fig. 1(b)). In addition, a group 1 HA-specific antibody (CR6261) did not react to either of the seal or human virus H3 proteins, which both belong to group 2 HA (Fig. 1(c)). These findings confirmed that the identity and purity of the recombinant HA proteins (Fig. 1(a)) produced by the baculovirus system were of the correct subtype.

Recombinant seal H3N8 HA binds to human and ferret lung tissues. In order to assess the ability of seal H3N8 virus to infect the upper respiratory tract of humans, which appears to be a requirement for efficient human-to-human transmission, we sought to investigate the binding patterns of its HA protein to physiological glycans present on human tissues. We compared the binding patterns of recombinant HA proteins derived from seal H3N8 against that of a seasonal human H3N2 virus (A/Wyoming/03/2003) to fixed human lung and tracheal tissue sections. HA proteins were allowed to bind to respiratory tissues, bound HA was detected by immuno-staining and the results were verified against negative control mock-stained tissue sections. Our results revealed that, in contrast to human H3N2, seal H3N8 HA exhibited minimal to no binding to the human trachea (Fig. 2). Nonetheless, it displayed greater binding affinity than the human H3N2 to human lung tissues. Quantifying the HA-specific signal revealed that human H3 bound efficiently to both human and ferret tracheas ($p = 0.0007$, 0.1250 respectively), whereas no binding signal could be quantified for the seal H3 on tracheal tissues. Both HA proteins showed binding to the cells lining the human and ferret lung alveolar tissues. Furthermore, the seal H3 bound stronger to human lungs than did the human H3, but the human H3 bound stronger to ferret lungs than did the seal H3 (Fig. 3, $p = 0.0007$, 0.0029 respectively).

Recombinant seal H3N8 HA binds to α-2,3 sialylated glycans. To obtain a more detailed and comprehensive picture of the receptor binding patterns of seal H3N8 HA, we tested its binding to an array of 610 glycans. Seal H3N8 virus harbored polymerase mutations that were indicative of mammalian adaptation[13], which raised some concerns that this virus could potentially infect and efficiently spread in humans or other mammalian species. Therefore, we compared the seal HA binding pattern to that of the human-adapted H3N2 strain. Two methods were used for testing HA binding, one where the protein sample, primary antibody and secondary antibody were added sequentially to the slide, and the second where all reagents were pre-complexed together in a tube before adding to the slide. A high degree of overlap was observed for the glycan hits detected by both methods (Fig. 4(a)). The pre-complexing method has shown improved binding and slightly elevated fluorescence signals (relative fluorescence units or RFU) than when reagents were added sequentially (Fig. 4(a)). The elevated signal obtained with the pre-complexing method could be due to the increased number of antibody linked binding sites, which would result in a higher avidity to bind to HA. Supplementary Tables S1–S4 show all glycan hits (with p-values of less than 0.01) identified in our study for both HAs tested. One glycan, Neu5Ac α2–3Galb1–4(Fuca1–3)(6S)GlcNAcb, abbreviated as 6-sulfo sialyl Lewis X (Su-SLe^x), was a common target for both HA proteins. Our glycan microarray screening (Fig. 4(a)) revealed that the H3N8 HA primarily binds to α-2,3 sialylated glycans similar to most avian adapted HAs[29]. Conversely, the human H3N2 HA binds predominantly to α-2,6 sialylated glycans, although some binding to α-2,3 sialylated glycans was also observed, which is consistent with previous studies[29].

Using a qualitative assessment of the glycan structural features shared among the statistically significant hits from the array (Fig. 4(b)), we found that the human-adapted HA showed preferences for α-2,6 sialosides with >2 Gal-GlcNAc extensions (herein denoted long α-2,6 sialosides). Additionally, several of the top hits contained α-1,3 fucosylated GlcNAc residues on the second or third GlcNAc (relative to the penultimate sialic acid). The presence of a fucosylated GlcNAc had a relatively minor impact on H3N2 binding. In contrast, the seal H3N8 HA largely bound α-2,3-linked sialosides, with only one Gal-GlcNAc repeat (herein denoted short α-2,3 sialosides). Some modifications, such as 6-O-sulfation and fucosylation, were observed branching off of the first GlcNAc (relative to the penultimate sialic acid), however, these had minor effects on HA binding.

Seal H3N8 viral growth kinetics. We examined the replication efficiency of seal H3N8, human H3N2 (A/Brisbane/10/2007) and avian H3N8 (A/American green-winged teal/Interior Alaska/10BM07649R0/2010) viruses in three types of cells: Madin-Darby Canine Kidney (MDCK) and human lung carcinoma cells (A549), and an avian cell line: duck embryo fibroblasts (DEF). Cell monolayers were infected in duplicate at a multiplicity of infection (MOI) of 1, and viral titers in the supernatants were monitored over a period of 72 hours (h). All three viruses replicated efficiently in MDCK cells, however human H3N2 virus titers were significantly higher than the seal and avian H3N8 that replicated to comparable levels, particularly at 24 and 48 h post-infection (Fig. 5(a)). A similar pattern was observed for the human H3N2 virus in A549 cells (Fig. 5(b)), where it exhibited titers that were about two orders of magnitude higher than the other two viruses. Interestingly, the seal H3N8 virus titers were significantly higher than that of the avian H3N8 virus at 24 and 48 h post-infection ($p = 0.0294$). The poorest replication kinetics for all three viruses were observed in DEF cells, where the seal H3N8 virus displayed significantly lower titers than its avian counterpart, particularly at 48 h ($p = 0.0286$) and 72 h ($p = 0.0265$) post-infection (Fig. 5(c)).

Discussion

In nature, influenza viruses inhabit the guts of wild birds primarily belonging to the orders Anseriformes and Charadriformes. In these birds, infection is generally clinically asymptomatic and the virus replicates mainly in the intestinal tract. Occasionally, these viruses acquire mutations that allow them to switch hosts and infect domestic birds and mammals, where they can replicate in the respiratory tract or other tissues, causing mild to severe disease symptoms[1]. This report is an in depth assessment of the receptor binding specificity of the seal H3N8 virus (A/harbour seal/New Hampshire/179629/2011) that emerged, most likely from avian origins, in the

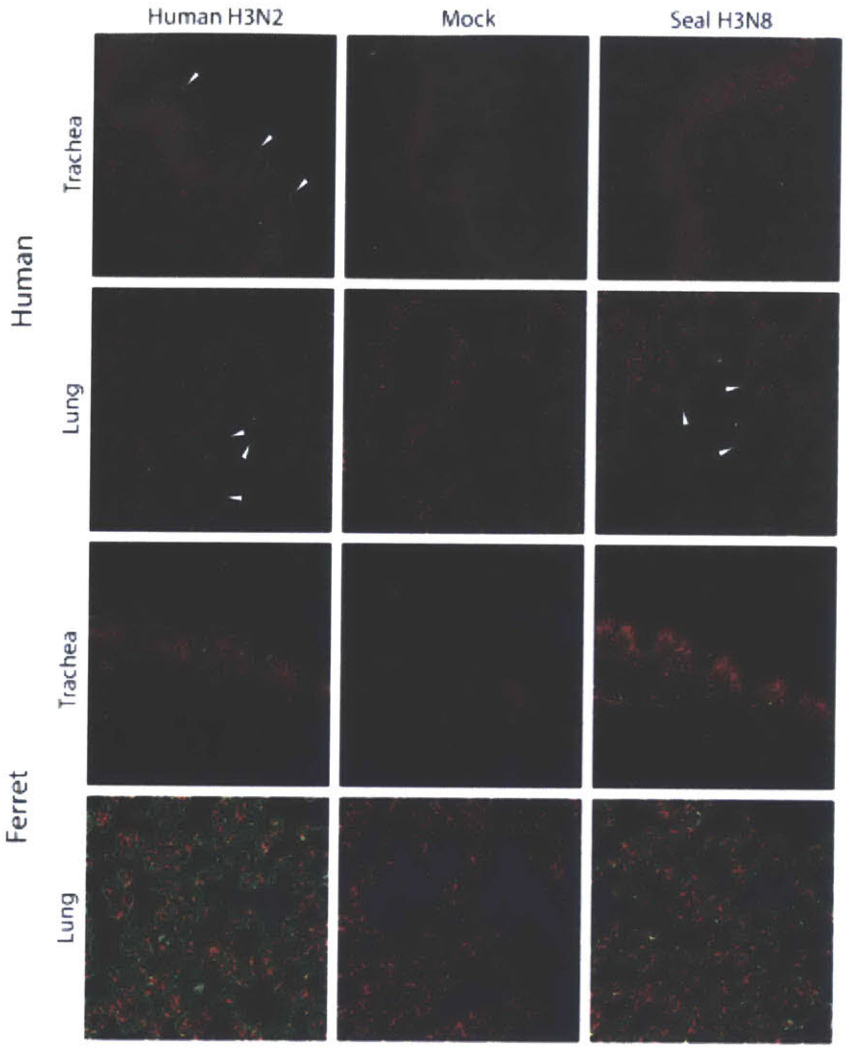


Figure 2. HA immunostaining of human respiratory tissues. Confocal microscopy images (20X) showing the varying binding affinity of human H3N2 and seal H3N8 HA proteins (green pointed by white arrows) to human and ferret tracheal and lung tissues (nuclei stained red).

New England harbor seal population in late 2011[17]. Based on the differential ability to agglutinate guinea pig and swine red blood cells and co-staining of viral HA and SAα-2,6-positive seal respiratory epithelium, Anthony *et al.* concluded that this virus was able to bind to both SAα-2,3 and SAα-2,6 receptors, a feature of avian viruses adapting to humans[17]. Another recent study reached similar conclusions based on a solid-phase binding assay that relied on only 2 types of biotinylated glycans (α-2,3'SL or α-2,6'SL)[30]. We therefore performed a series of experiments to assess in greater detail the ability of a recombinant HA of the seal H3N8 virus to bind to physiological glycans present on human and ferret tissues and to a large representative library of 610 chemically synthesized glycans on an array format. Previous studies have shown there is no difference in receptor binding specificity of recombinant HA proteins expressed in mammalian and insect cells[31,32]. Moreover, the physiological human respiratory glycans were shown to be well represented on the array produced by the Consortium of Functional Glycomics used in our experiments[18]. Therefore, we are confident that our recombinant HA binding studies reflects the behavior of a native viral HA. Contrary to an earlier study[30], our data suggested that seal H3N8 recombinant HA has retained its avian receptor binding specificity as physiologically relevant immunohistochemistry

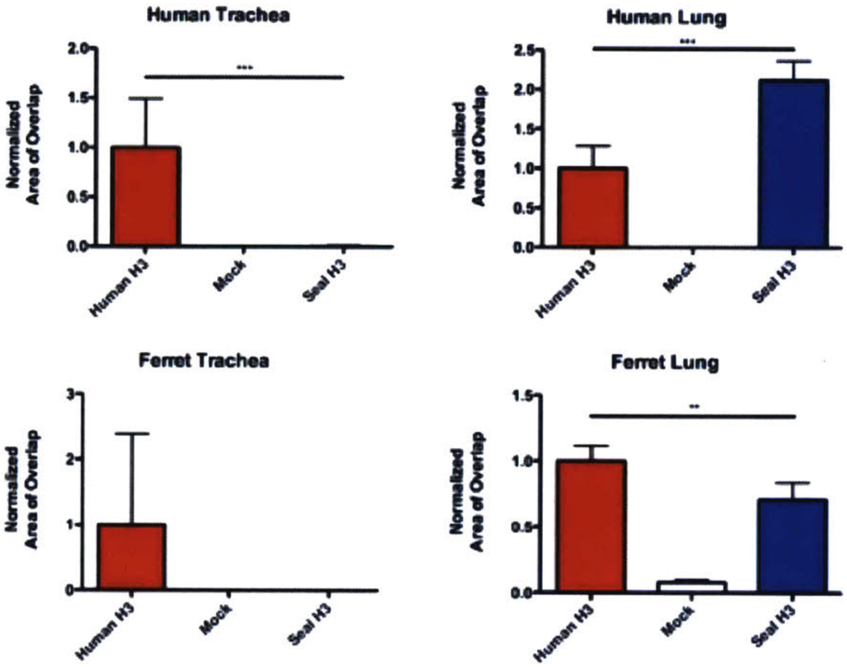


Figure 3. Quantitation of recombinant HA binding to human and ferret respiratory tissues. A minimum of three images were processed for each measurement. Error bars represent the 95% confidence intervals. Statistically significant differences between the human and seal H3 staining patterns across tested tissues are denoted by horizontal lines and asterisks.

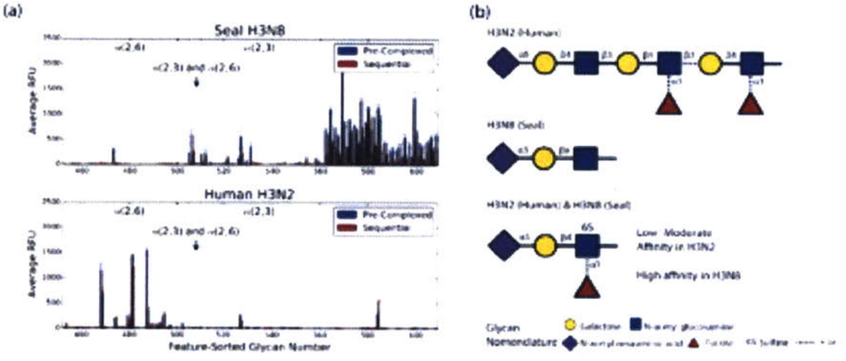


Figure 4. Glycan binding profiles of human H3N2 and seal H3N8 recombinant HA proteins. (a) The glycans on our array were sorted according to the type of their sialic acid linkages (X-axis) and plotted against the averaged relative fluorescence unit (RFU) values (Y-axis). Error bars represent SEM of 4 RFU values for each glycan tested in our array. **(b)** Glycan cartoons representing the most prevalent motif bound by either H3N2 HA (top), H3N8 HA (middle) or both H3N2 and H3N8 HA (bottom). Dotted lines indicate mixed presence and absence among commonly bound glycan motifs.

demonstrated binding to human lung, but not the tracheal tissues, and as seal H3N8 recombinant HA showed strong preference for SAα-2,3Gal receptors on a mammalian glycan array (Figs 2 and 3). These findings are in agreement with the recently published HA binding data that relied on an array composed of a smaller group of 96

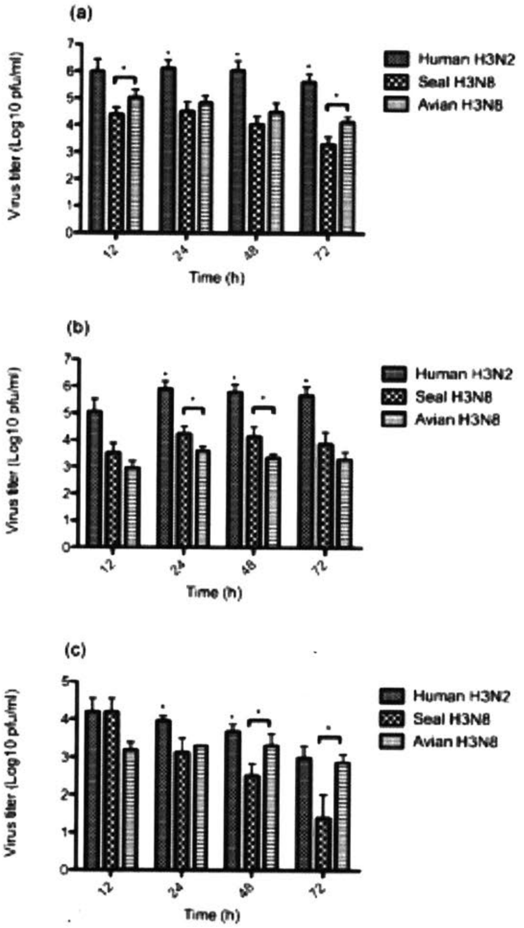


Figure 5. Replication kinetics of human H3N2, seal H3N8 and avian H3N8 viruses in MDCK (**a**), A549 (**b**) and DEF (**c**) cells. Cell monolayers seeded in 12-well plates were infected in duplicate, then allowed to adsorb for 60 minutes, then the virus inoculum were aspirated and cells were washed with PBS. Viral supernatants were collected at 12, 24, 48 and 72 hours post-infection and titrated by plaque assay in MDCK cells. Error bars represent the 95% confidence intervals (CI) of the mean of two independent experiments. Human H3N2 virus titers that were significantly higher than both of its seal and avian H3N8 counterparts are denoted by an asterisk (*). Statistically significant differences between the seal and avian H3N8 viruses are denoted by a horizontal line and an asterisk at the specified time points.

glycans[33] in which the binding patterns of a whole seal H3N8 virus were generally consistent with the recombinant HA protein used in our experiment. As expected, the HA of our human-adapted H3N2 (A/Wyoming/03/03) positive control showed strong binding to human trachea and moderate binding to alveolar tissues, which is consistent with previously published tissue staining data[26,34]. We also observed a similar staining pattern in ferret respiratory tissues, where the human H3N2 HA, unlike its seal H3N8 counterpart, bound to ferret tracheal tissue sections. This is consistent with the previous observations that the glycan receptor distribution on the human and ferret respiratory tissues is similar[35,36], though not identical[37].

Although it is widely accepted that human-adapted and avian strains prefer SAα-2,6 receptors and SAα-2,3Gal respectively, this correlation is not absolute for all influenza virus subtypes[38]. Binding to SAα-2,3Gal receptors did not restrict the avian-to-human transmission of human H5N1 (A/HK/156/97) viruses isolated from the 1997 Hong Kong outbreak[39]. Glycan array analysis of the highly pathogenic H5N1 (A/Vietnam/1203/2004) virus, which was isolated from a Vietnamese bird flu victim, revealed a binding preference for sialylated glycans with SAα-2,3 linkage[40]. Clinically, H5N1 infection in humans is characterized by lung involvement, where virus replication mainly takes place[41]. Furthermore, immuno-staining studies of an H5N1 virus (A/Vietnam/1194/04) revealed strong binding to type II pneumocytes of human lungs[42]. These findings indicated that highly pathogenic IAVs retaining avian receptor binding affinity could replicate and cause fatal disease in humans without a significant change in its receptor-binding affinity. Here we show that the seal H3N8 virus replicated more efficiently in human lung A549 carcinoma cells than its avian counterpart (Fig. 5(b)), suggesting that it could potentially establish infection in the lower respiratory tract of humans. However, since alterations of receptor binding preference seem to be a prerequisite for efficient human-to-human transmission[18,43], the lack of α-2,6 sialylated glycan binding by seal H3N8 HA indicates that this virus has not yet acquired the mutations required for human adaptation and is unlikely to spread efficiently among humans. In their study of seal H3N8 virus airborne transmission, Karlsson and colleagues have detected an HA A134T mutation, known to alter the receptor-binding specificity of avian H5N1 viruses from α-2,3 to α-2,6 sialylated glycans, in viruses recovered from aerosol/droplet contact ferrets[30]. However, it is not clear whether this mutation facilitated aerosol transmission or if it emerged in the sentinel ferrets after transmission. In either case, combining with our results raises a question of whether the naturally acquired polymerase mutations (e.g. PB2 D701N) in the seal H3N8 virus may have more of an impact on ferret transmissibility than previously understood.

Additional features of cell surface glycans beyond the terminal sialic acid linkage were shown to be important in the binding of human-adapted versus avian-adapted HAs to their respective glycan receptors[44]. The breadth of the array used in our experiments, which is composed of 610 glycans, enabled us to also probe the structural determinants of the human versus seal H3 receptor binding. The long α-2,6 motif was found to be the most critical determinant of human-adapted HA binding. On the other hand, the seal H3N8 HA bound mainly to short α-2,3-linked sialosides, with only one Gal-GlcNAc repeat (Fig. 4(b)). These findings are supported by previous studies demonstrating that receptor topology governs specificity of human and avian receptors. The long α-2,6 sialosides are capable of adopting a flexible umbrella-like topology, and are the predominant receptor type for pandemic human-adapted HA. The short α-2,3 sialosides have been shown to adopt a cone-like topology, and are the predominant receptor type for avian-adapted HA[24]. Thus, we believe that the seal H3N8 shares receptor-binding characteristics similar to those of avian-adapted HAs. Interestingly, the fact that we could detect binding of both human and seal H3s to Su-SLe^{x} indicates that the seal H3N8 virus could be diverging away from its avian ancestors[17]. Enhanced binding to sulfated and/or fucosylated glycans with α-2,3 linkages, particularly Su-SLe^{x}, was a common feature of IAVs isolated from terrestrial poultry, pigs and horses, but not duck viruses[45]. An earlier glycan array study has also shown that, in contrast to a duck H3 virus (A/Duck/Ukraine/1963), the HA of several human H1 and H3 viruses bound to Su-SLe^{x} [29].

In conclusion, seal H3N8 virus still maintains the avian-type receptor specificity, binds to human lung tissues and replicates in human lung carcinoma cells, which raises concerns about its potential to establish infection in the lower respiratory tract of humans. However, we believe that certain additional mutations will be required for this virus to gain human transmissibility. Data presented in this study coupled with the recently published seal H3N8 HA crystal structure[33], could provide impetus for future studies using similar approaches to unravel mutations that could potentially facilitate binding to human receptors. This study also helps clarify our understanding of the circulation and adaptation of influenza virus in seals, which is needed for early detection and characterization of viruses with an enhanced potential to infect humans and to evaluate if marine mammal populations could be a reservoir for mammalian adaptation of potentially pandemic human influenza virus.

Methods

Viruses, cells and tissues. Seal H3N8 (A/harbour seal/New Hampshire/179629/2011) was obtained from Dr. Hon Ip (National Wildlife Health Center, Madison Wisconsin). Avian H3N8 (A/American green-winged teal/Interior Alaska/10BM07649R0/2010) was one of our own Alaskan isolates. Human H3N2 (A/Brisbane/10/2007) was obtained from Biodefense and Emerging Infections Research Resources Repository (BEI). Viral stocks were prepared by inoculating 10-day old embryonated chicken eggs and harvesting the allantoic fluid 3 days later. A549 lung carcinoma cells (ATCC CCL-185), DEF duck embryo fibroblasts (ATCC CCL-141) and MDCK cells (ATCC CCL-34) were maintained in DMEM containing 10% FBS and penicillin/streptomycin at a final concentration of 50 IU/ml penicillin and 50 μg/ml streptomycin. Formalin-fixed paraffin-embedded tissue sections of the human trachea and lung were purchased from BioChain. Archival normal ferret tissue specimens were kindly provided by the Histology Laboratory of MIT's Division of Comparative Medicine. These tissues were fixed in 10% neutral buffered formalin for 24 hours and processed by routine paraffin embedding and sectioned at 4–6 μm for subsequent immuno-staining.

Viral growth kinetics and titration. Twelve-well plates were seeded with MDCK, A549 or DEF cells and allowed to grow until confluent monolayers were obtained. On the day of the experiment, one monolayer from each type of cells was trypsinized in 0.25% Trypsin-EDTA and counted in a hemocytometer. Titrated viral stocks were diluted and used to infect cells in duplicates at a multiplicity of infection (MOI) of 1. Briefly, viruses were allowed to adsorb for 60 minutes (mn), then the virus inoculum were aspirated and cells were washed once with sterile PBS. Supernatants were collected at 0, 12, 24, 48 and 72 hours post-infection and titrated by plaque assay on fresh MDCK monolayers.

Recombinant HA expression and purification. Recombinant HAs were expressed as described before[46]. Briefly, genes encoding the ectodomains of the A/Wyoming/03/03 and A/harbour seal/New Hampshire/179629/2011 HAs were cloned into a modified pFastBacDual (Invitrogen) baculovirus transfer vector that harbors a C-terminal T4 foldon trimerization domain, a thrombin cleavage site and a hexahistidine tag. The identity of the recombinant baculovirus vectors was verified by Sanger sequencing, which was carried out by the sequencing services of Macrogen. To generate recombinant bacmids, the transfer plasmids were transformed into

DH10Bac competent bacteria (Invitrogen). Bacmids were then transfected into Sf9 insect cells to generate recombinant baculovirus. Cell supernatants were incubated with NiNTA resin (Qiagen) and protein preps were concentrated and buffer exchanged to pH 7.4 PBS using Amicon Ultra centrifugation columns (Millipore). Recombinant HA proteins were checked for structural integrity and identity using SDS-PAGE and ELISA as described before[47].

Immuno-staining of human and ferret respiratory tissues. Tissue staining was carried out as previously described[48]. Briefly, the paraffin coating was melted, and slides were then blocked with 1% BSA-PBS, followed by incubation with HA pre-complexes at a ratio of 4:2:1 [HA (seal H3N8, human H3N2, or mock): primary antibody (mouse anti-His from Abcam): Secondary antibody (Goat anti-mouse labeled with Alexa Fluor from Lifetech)]. Slides were then immersed in propidium iodide (Lifetech) at a final concentration 1:100, then washed and finally mounted in anti-fade reagent (Lifetech) for confocal imaging using Zeiss 700 laser scanning microscope.

Image quantitation. We used the scikit-image Python package for image quantification[49]. Briefly, images of human H3 binding trachea were treated as positive controls, and mock-stained slides were treated as negative controls. The images were separated into their red and green channels. In our particular staining protocol, the nuclei, which represent cells, are not directly in contact with the HA protein. Therefore, instead of computing the amount of red-green overlap, we sought to quantify the amount of green (protein) associated within an area around the red (cells) or vice versa (Figs S1 and S2). To identify regions of significant red (nuclei) or green (HA protein), we first applied an intensity threshold computed based on the positive control images, by using Otsu's method[50]. We then sought to delineate a region around the nuclei or the HA protein boundaries. This was accomplished by computing the entropy of the thresholded red and green channels within a radius of 15 and 10 pixels respectively, and then thresholding the resulting image, identifying regions of significant entropy. A demonstration of this procedure is provided as an IPython HTML notebook on Github (supplementary materials). We then computed the number of pixels overlapping between the nuclei boundary regions and the HA protein regions. The threshold values computed for the positive control slides were averaged, and this value was also used for the negative control and seal H3 samples.

Glycan array screening. Receptor binding specificities of seal H3N8 and human H3N2 HA recombinant proteins were tested on an array comprising 610 glycan targets. A list of the glycans used in this study (array version 5.1) can be found here: http://www.functionalglycomics.org/static/consortium/resources/resourcecoreh8.shtml. This array was manufactured by the Consortium for Functional Glycomics (CFG)[51]. Each protein was tested in 6 replicates at a concentration of 200 μg/ml in a binding buffer (20 mM Tris-HCL pH 7.4, 150 mM sodium chloride, 2 mM calcium chloride, 2 mM magnesium chloride, 0.05% Tween 20 and 1% BSA). HA binding was tested by two methods, one where the protein sample, mouse anti-His primary antibody (Abcam) and Alexa-labeled anti-mouse secondary antibody (provided by CFG) were added in sequential steps to the slide, and another where all reagents were pre-complexed in a tube before adding to the slide. Each protein was tested in 6 replicates. The highest and lowest point from each set of 6 replicates was excluded to eliminate some of the false hits. The remaining 4 relative fluorescence unit (RFU) values were averaged and plotted for each protein (Fig. 4(a)). To identify the preferred receptor-binding motif of the seal and human subtype 3 hemagglutinin proteins (H3s), we analyzed the 'hits' identified on the glycan array. We assessed four key features: i) terminal sialic acid linkage (α-2,3 or α-2,6), ii) number of Hex-HexNAc repeats ($n = 1$ or $n > 1$; predominantly Gal-GlcNAc), iii) sulfation, and iv) fucosylation. These features were chosen because they were previously identified as determinants of hemagglutinin binding in human or avian adapted viruses[24,43]. In Tables S1–S4, the presence of a feature was indicated with a one and the absence with a zero. In certain cases, such as $(\text{Hex-HexNAc})_n$ repeats where $n > 1$, the exact number of repeats (n) was indicated in brackets next to the 1. In cases where two or more structurally identical branches were attached to the N-linked glycan 'core' ($\text{Man}_3\text{GlcNAc}_2$) or a GalNAc core, the number of branches was noted but only one branch was taken into account when assessing structural features. A qualitative assessment of the features was performed across the top binders, and a representative cartoon was drawn to depict common features of the glycans that bound each HA (Fig. 4(b)).

Statistical analysis. For the glycan array experiments, we used the data to estimate a baseline value for non-binders. As we expect the integer values of the RFU to be distributed continuously at the non-binding baseline, and positive hits to be "broken" off from this continuity, we took the value after the first "break" in continuity as the estimated value for non-binding baseline. This was done as opposed to picking an arbitrary value to use across all data sets, in order to account for variability between each experiment. We computed a t-score, under the null hypothesis of non-binding using the estimated non-binding baseline value, and computed the corresponding p-value using a one-tailed t-test with 3 degrees of freedom. We then selected hits that had a p-value of less than 0.01 (supplementary Tables 1–4). For image quantitation, the Mann-Whitney U-test was used for comparisons between the human H3 and seal H3 on the lung and tracheal tissues of human and ferret (Fig. 3). Because the seal H3 binding values were all zero on the ferret trachea, the Wilcoxon signed rank test was used instead. For virus replication kinetics, the paired t-test was used for reported comparisons (Fig. 5).

References

1. Taubenberger, J. K. & Kash, J. C. Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* 7, 440–451, doi: 10.1016/j.chom.2010.05.009 (2010).
2. Runstadler, J., Hill, N., Hussein, I. T., Puryear, W. & Keogh, M. Connecting the study of wild influenza with the potential for pandemic disease. *Infect Genet Evol* 17, 162–187, doi: 10.1016/j.meegid.2013.02.020 (2013).
3. Fereidouni, S., Munoz, O., Von Dobschuetz, S. & De Nardi, M. Influenza Virus Infection of Marine Mammals. *Ecohealth*, doi: 10.1007/s10393-014-0968-1 (2014).

4. Webster, R. G., Geraci, J., Petursson, G. & Skirnisson, K. Conjunctivitis in human beings caused by influenza A virus of seals. *N Engl J Med* **304**, 911, doi: 10.1056/NEJM198104093041515 (1981).
5. Hinshaw, V. S. *et al.* Are seals frequently infected with avian influenza viruses? *J Virol* **51**, 863–865 (1984).
6. Callan, R. J., Early, G., Kida, H. & Hinshaw, V. S. The appearance of H3 influenza viruses in seals. *J Gen Virol* **76** (Pt 1), 199–203 (1995).
7. Goldstein, T. *et al.* Pandemic H1N1 influenza isolated from free-ranging Northern Elephant Seals in 2010 off the central California coast. *PLoS One* **8**, e62259, doi: 10.1371/journal.pone.0062259 (2013).
8. Zohari, S., Neimanis, A., Harkonen, T., Moraeus, C. & Valarcher, J. Avian influenza A(H10N7) virus involvement in mass mortality of harbour seals (Phoca vitulina) in Sweden, March through October 2014. *Euro Surveill* **19** (2014).
9. Nielsen, O., Clavijo, A. & Boughen, J. A. Serologic evidence of influenza A infection in marine mammals of arctic Canada. *J Wildl Dis* **37**, 820–825, doi: 10.7589/0090-3558-37.4.820 (2001).
10. Ohishi, K. *et al.* Serological evidence of transmission of human influenza A and B viruses to Caspian seals (Phoca caspica). *Microbiol Immunol* **46**, 639–644 (2002).
11. Ohishi, K. *et al.* Antibodies to human-related H3 influenza A virus in Baikal seals (Phoca sibirica) and ringed seals (Phoca hispida) in Russia. *Microbiol Immunol* **48**, 905–909 (2004).
12. Fujii, K. *et al.* Serological evidence of influenza A virus infection in Kuril harbor seals (Phoca vitulina stejnegeri) of Hokkaido, Japan. *J Vet Med Sci* **69**, 259–263 (2007).
13. Murphy, B. R. *et al.* Evaluation of the A/Seal/Mass/1/80 virus in squirrel monkeys. *Infect Immun* **42**, 424–426 (1983).
14. White, V. C. A review of influenza viruses in seals and the implications for public health. *US Army Med Dep J*, 45–50 (2013).
15. Bowen, W. D., McMillan, J. & Mohn, R. Sustained exponential population growth of grey seals at Sable Island, Nova Scotia. **60**, 1265–1274 (2003).
16. Waltzek, T. B., Cortes-Hinojosa, G., Wellehan, J. F., Jr. & Gray, G. C. Marine mammal zoonoses: a review of disease manifestations. *Zoonoses Public Health* **59**, 521–535, doi: 10.1111/j.1863-2378.2012.01492.x (2012).
17. Anthony, S. J. *et al.* Emergence of fatal avian influenza in New England harbor seals. *MBio* **3**, e00166–00112, doi: 10.1128/mBio.00166-12 (2012).
18. de Graaf, M. & Fouchier, R. A. Role of receptor binding specificity in influenza A virus transmission and pathogenesis. *EMBO J* **33**, 823–841, doi: 10.1002/embj.201387442 (2014).
19. Shi, Y., Wu, Y., Zhang, W., Qi, J. & Gao, G. F. Enabling the 'host jump': structural determinants of receptor-binding specificity in influenza A viruses. *Nat Rev Microbiol* **12**, 822–831, doi: 10.1038/nrmicro3362 (2014).
20. Nobusawa, E. *et al.* Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology* **182**, 475–485 (1991).
21. Gambaryan, A. S. *et al.* Specification of receptor-binding phenotypes of influenza virus isolates from different hosts using synthetic sialylglycopolymers: non-egg-adapted human H1 and H3 influenza A and influenza B viruses share a common high binding affinity for 6'-sialyl(N-acetyllactosamine). *Virology* **232**, 345–350, doi: 10.1006/viro.1997.8572 (1997).
22. Matrosovich, M. *et al.* Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *J Virol* **74**, 8502–8512 (2000).
23. Glaser, L. *et al.* A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J Virol* **79**, 11533–11536, doi: 10.1128/JVI.79.17.11533-11536.2005 (2005).
24. Chandrasekaran, A. *et al.* Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin. *Nat Biotechnol* **26**, 107–113, doi: 10.1038/nbt1375 (2008).
25. Stevens, J. *et al.* Receptor specificity of influenza A H3N2 viruses isolated in mammalian cells and embryonated chicken eggs. *J Virol* **84**, 8287–8299, doi: 10.1128/JVI.00058-10 (2010).
26. Shinya, K. *et al.* Avian flu: influenza virus receptors in the human airway. *Nature* **440**, 435–436, doi: 10.1038/440435a (2006).
27. Stevens, J. *et al.* Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870, doi: 10.1126/science.1093373 (2004).
28. Ekiert, D. C. *et al.* Antibody recognition of a highly conserved influenza virus epitope. *Science* **324**, 246–251, doi: 10.1126/science.1171491 (2009).
29. Stevens, J. *et al.* Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *J Mol Biol* **355**, 1143–1155, doi: 10.1016/j.jmb.2005.11.002 (2006).
30. Karlsson, E. A. *et al.* Respiratory transmission of an avian H3N8 influenza virus isolated from a harbour seal. *Nat Commun* **5**, 4791, doi: 10.1038/ncomms5791 (2014).
31. Ramos, I. *et al.* H7N9 influenza viruses interact preferentially with alpha2,3-linked sialic acids and bind weakly to alpha2,6-linked sialic acids. *J Gen Virol* **94**, 2417–2423, doi: 10.1099/vir.0.056184-0 (2013).
32. Xu, R. *et al.* Preferential recognition of avian-like receptors in human influenza A H7N9 viruses. *Science* **342**, 1230–1235, doi: 10.1126/science.1243761 (2013).
33. Yang, H. *et al.* Structural and Functional Analysis of Surface Proteins from an A(H3N8) Influenza Virus Isolated from New England Harbor Seals. *J Virol* **89**, 2801–2812, doi: 10.1128/JVI.02723-14 (2015).
34. van Riel, D. *et al.* Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals. *Am J Pathol* **171**, 1215–1223, doi: 10.2353/ajpath.2007.070248 (2007).
35. Leigh, M. W., Connor, R. J., Kelm, S., Baum, L. G. & Paulson, J. C. Receptor specificity of influenza virus influences severity of illness in ferrets. *Vaccine* **13**, 1468–1473 (1995).
36. Kirkeby, S., Martel, C. J. & Aasted, B. Infection with human H1N1 influenza virus affects the expression of sialic acids of metaplastic mucous cells in the ferret airways. *Virus Res* **144**, 225–232, doi: 10.1016/j.virusres.2009.05.004 (2009).
37. Jia, N. *et al.* Glycomic characterization of respiratory tract tissues of ferrets: implications for its use in influenza virus infection studies. *J Biol Chem* **289**, 28489–28504, doi: 10.1074/jbc.M114.588541 (2014).
38. Zhang, H. Tissue and host tropism of influenza viruses: importance of quantitative analysis. *Sci China C Life Sci* **52**, 1101–1110, doi: 10.1007/s11427-009-0161-x (2009).
39. Matrosovich, M., Zhou, N., Kawaoka, Y. & Webster, R. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J Virol* **73**, 1146–1155 (1999).
40. Stevens, J. *et al.* Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* **312**, 404–410, doi: 10.1126/science.1124513 (2006).
41. Uiprasertkul, M. *et al.* Influenza A H5N1 replication sites in humans. *Emerg Infect Dis* **11**, 1036–1041, doi: 10.3201/eid1107.041313 (2005).
42. van Riel, D. *et al.* H5N1 Virus Attachment to Lower Respiratory Tract. *Science* **312**, 399, doi: 10.1126/science.1125548 (2006).
43. Wilks, S., de Graaf, M., Smith, D. J. & Burke, D. F. A review of influenza haemagglutinin receptor binding as it relates to pandemic properties. *Vaccine* **30**, 4369–4376, doi: 10.1016/j.vaccine.2012.02.076 (2012).
44. Viswanathan, K. *et al.* Glycans as receptors for influenza pathogenesis. *Glycoconj J* **27**, 561–570, doi: 10.1007/s10719-010-9303-4 (2010).
45. Gambaryan, A. S. *et al.* 6-sulfo sialyl Lewis X is the common receptor determinant recognized by H5, H6, H7 and H9 influenza viruses of terrestrial poultry. *Virol J* **5**, 85, doi: 10.1186/1743-422X-5-85 (2008).

46. Margine, I., Palese, P. & Krammer, F. Expression of Functional Recombinant Hemagglutinin and Neuraminidase Proteins from the Novel H7N9 Influenza Virus Using the Baculovirus Expression System. e51112, doi: 10.3791/51112 (2013).

47. Krammer, F. *et al.* Divergent H7 immunogens offer protection from H7N9 virus challenge. *J Virol* **88**, 3976–3985, doi: 10.1128/JVI.03095-13 (2014).

48. Tharakaraman, K. *et al.* Glycan receptor binding of the influenza A virus H7N9 hemagglutinin. *Cell* **153**, 1486–1493, doi: 10.1016/j.cell.2013.05.034 (2013).

49. van der Walt, S. *et al.* Scikit-image: Image processing in Python. Report No. 2167–9843 (PeerJ PrePrints, 2014).

50. Otsu, N. A threshold selection method from gray-level histograms. **9**, 62–66, doi: citeulike-article-id:1116982 (1979).

51. Heimburg-Molinaro, J., Song, X., Smith, D. F. & Cummings, R. D. Preparation and analysis of glycan microarrays. *Curr Protoc Protein Sci* Chapter 12, Unit12 10, doi: 10.1002/0471140864.ps1210s64 (2011).

Acknowledgements

The authors would like to acknowledge the Centers of Excellence for Influenza Research and Surveillance (CEIRS), Center for Research on Influenza Pathogenesis (NIAID HHSN266200700010C) for funding. This publication is partially the result of research sponsored by The MIT Sea Grant College Program, under NOAA Grant Number, MIT SG Project Number 2013-DOH-45-LEV. This work could not be made possible without support from the Consortium for Functional Glycomics, funded by the NIGMS (GM62116 and GM98791) for services provided by the Glycan Array Synthesis Core (The Scripps Research Institute, LaJolla, CA) that produced the mammalian glycan microarray and the Protein-Glycan Interaction Core (Emory University School of Medicine, Atlanta, GA) that assisted with analysis of samples on the array. We would like also to thank Hon Ip, from the USGS National Wildlife Health Center, Madison, for providing us with the A/harbour seal/New Hampshire/179629/2011 virus. Finally, we want to thank the W.M. Keck Microscopy Facility of the Whitehead Institute for their imaging support, James G. Fox, Suresh Muthupalani and the Histology Laboratory of the Division of Comparative Medicine at MIT for providing the ferret tissue sections.

Author Contributions

I.H., M.E., E.M. and J.R. conceived and designed the experiments. I.H. and M.E. conducted the experiments. F.K. made the recombinant HA proteins. E.M. conducted the computational analysis. K.V., N.S., D.Q. and R.S. provided technical and analytical support for the tissue staining and glycan array experiments. I.H., E.M. and J.R. analyzed the experimental data and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hussein, I. T. M. *et al.* New England harbor seal H3N8 influenza virus retains avian-like receptor specificity. *Sci. Rep.* **6**, 21428; doi: 10.1038/srep21428 (2016).

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators. My contribution included: Interpretation of the glycan array binding data to determine the glycan binding specificity, as well as technical and analytical support for assessing HA-glycan interactions. I'd like to thank Eric Ma, Islam Hussein, Florian Krammer, Michael Estrin, Devin Quinlan, Karthik Viswanathan, Ram Sasisekharan and Johnathan Runstadler for their contributions to this work. I would like to acknowledge the following agencies who provided funding to support this work: the Centers of Excellence for Influenza Research and Surveillance (CEIRS), Center for Research on Influenza Pathogenesis (NIAID HHSN266200700010C), and the National Institutes of Health (T32 ES007020-38).

Chapter 7 : Development of Integrated Approaches to Study the Glycome

Summary and Significance

Glycans are abundant on the cell surface and at the cell-ECM interface where they mediate interactions between cells and their microenvironment. Given the growing interest in understanding how cell-microenvironment interactions modulate disease pathophysiology, there is a great need to develop, implement and apply approaches to study cell surface glycans. Glycans are expressed as a heterogeneous ensemble of structures where they interact with glycan binding proteins to mediate biological processes. To understand glycan function, a first key step is to characterize the structures present in the glycan ensemble (known as the glycome). However, due to glycan structural complexity there is no single tool capable of capturing all these features and multiple analytical tools and methods are needed. Importantly, glycomics measurements enable the generation of hypothesis through correlation with biological phenotypes. Testing of these hypotheses requires the ability to manipulate the glycome using molecular biology methods (i.e. knockout of biosynthetic genes) to build functional relationships. Thus, a robust glycomics approach must incorporate methods to characterize the structural composition of the glycome alongside methods to understand the relationships between the glycome and the glycan biosynthesis genes. Ultimately, using this integrated approach (i.e. a functional glycomics approach) it is possible to converge on structure-function relationships.

In this Chapter, I implement an efficient integrated analytical approach to profile the glycome of cells, namely N- and O- linked glycans. The analytical workflow developed here include: i) MALDI-TOF/TOF analysis of permethylated glycan, ii) a solid phase lectin array analysis, and iii) transcriptomics analysis of glycogene expression. The analytical assays implemented, optimized and validated using model glycoproteins (data not shown). Finally, the integrated workflow was applied to characterize the N- and O-linked glycome associated with metastasis in a cell model of lung adenocarcinoma.

The approach developed here identified changes to the glycome at multiple levels (i.e. changes in abundance of specific glycan and changes to glycan motifs) and could correlate these changes to the glycan biosynthesis genes responsible. Together, these methods provide a robust and efficient way to characterize glycome and provide a key first step towards decoding structure-function relationships of glycans. Furthermore, although this section focused on developing an approach to study N- and O-linked glycans, the framework, expertise and insights developed here are generalizable to the study of other glycans such as HSGAGs.

Introduction

The characterization of glycan structures is central to uncovering their biological role. However, even with modern analytical tools, the unique chemical complexities of glycans make their analysis challenging. Many of the monosaccharide building blocks are isobaric stereoisomers (e.g. Glc, Man, and Gal are same MW), they can be branched or linear, and have varied anomeric configuration (α or β) and diverse linkage types between monosaccharides [2]. Furthermore, glycans can be neutral or acidic, variable in polymer size ($n=1\text{-}10^3$), and contain modifications like acetylation, phosphorylation, or sulfation [32]. These unique features give rise to a large number of isobaric isomers and a glycan pool with heterogeneous physicochemical properties. Considering these features, no single method is capable of measuring all of the features of the glycan structure present in the glycan pool. For example, mass-based characterization of glycans is not straightforward because mass spectrometry (MS) cannot readily distinguish isobaric structures, nor provide definitive details on anomericity or glycosidic linkage. Conversely, liquid chromatography can resolve isomers but cannot provide direct structural information, thus it must be coupled with sequential enzyme digests or complemented by direct mass composition analysis. These examples illustrate the need for multidimensional methods for the characterization and integration across techniques to probe the glycome.

As multi-methodological analysis is required, there is a trade-off between measuring detailed structural information and high-throughput compositional analysis (which yields valuable, yet incomplete structural information). For instance, glycosylation changes that occur during cancer have been the focus of many diagnostic or clinical biomarker discovery efforts. For these studies, the profiling of alterations in structural features (i.e. changes in core fucosylation or terminal sialylation [215]) or quantifying the relative amounts of certain structures (i.e. abundance of fucosylated triantennary glycan in hepatocellular carcinoma [216]) are informative. In contrast, studies concerned with *de novo* characterization of novel glycan structure [217] or mechanistic determination of glycan-protein interactions require rigorous determination of composition, linkage, and anomericity to determine exact sequence or 3D structural conformation [91,92].

A number of techniques, such as mass spectrometry (MS), high performance liquid chromatography (HPLC), capillary electrophoresis (CE), nuclear magnetic resonance (NMR), and glycan binding/degrading reagents are used together to identify structure. Each method offers a unique and complementary dimension of data, bringing with it a unique set of advantages and limitations.

In the context of this thesis, the objective was to implement an efficient integrated approach to profile the cellular glycome. The development of an appropriate analytical approach was divided into three stages: First, a comprehensive literature review on analytical strategies for glycomics analysis was performed to identify the collection of analytical tools that might be suitable for our analysis. Next, select analytical assays were implemented and validated using model glycoproteins (data not shown). Finally, the integrated analysis was applied to characterize the N- and O-linked glycome associated with metastasis in a cell model of lung adenocarcinoma.

Based on the literature review, an integrated analytical approach comprised of three complementary methods was implemented. These methods include: i) MALDI-TOF/TOF analysis of permethylated glycan, ii) a solid phase lectin array analysis, and iii) transcriptomics analysis of glycogene expression. The next three sections detail the rationale for selecting these tools, the types of structural information obtained using these technologies, and the special considerations for applying these tools to study glycans.

MALDI-TOF/TOF analysis of permethylated glycan

Similar to proteomics and metabolomics, mass spectrometry (MS) has become the primary workhorse for glycomics analyses. Mass-based analysis of glycan is advantageous compared to indirect analysis methods (e.g. liquid chromatography migration time), due to the unambiguous structural information obtained. MS is used for glycan compositional profiling, while MS/MS or MS^n can elucidate detailed structural information on positional and linkage isomers. Notably, MS has not superseded the need for other methods (e.g. NMR, exoglycosidase treatment, or lectin analysis), and orthogonal means of characterization are still used to further inform and validate structural data obtained using MS.

Importantly, a key first step release of the glycan pool from the glycoconjugate carrier is a critical step in structural glycomics characterization. Due to the differences in the monosaccharide linkage to the peptide backbone, N-, and O-linked each required a unique method of release. N-glycans share a common core structure ($Man_3GlcNAc_2$), and are released using Peptide-N-glycosidase F (PNGaseF), an enzyme that cleaves the bond between the peptide asparagine and the N-Glycan's core GlcNAc [218]. Chemical methods for N-glycan release such as hydrazinolysis and β-elimination have also been developed but are not favorable due to their use of toxic reagents [219]. In contrast, Ser/Thr-linked glycans are more diverse in their attachment to the protein, and there is no single enzyme that is capable of releasing all O-glycans. For mucin-type O-glycans, chemical methods such as hydrazinolysis [219] and β-elimination [220] are used. Alkaline β-elimination cleaves the O-Glycosidic bond and releases the O-Glycans in the reduced alditol form [220]. In the context of the method developed here, N-Glycans are released enzymatically using PNGaseF and O-linked glycans are released using Alkaline β-elimination.

Selection of a glycan preparation and ionization/detection method is also a critical step in MS methods development. While there are several ionization methods that have been used for glycomics analysis, Matrix-assisted laser desorption/ionization (MALDI) is the most widely used due to the ease of sample preparation, potential for automation, speed of data acquisition, and the simplicity of data interpretation (i.e. only singly charged ions observed) [221]. MALDI can be used to profile native glycans, however, one major limitation of this technique is that the acidic moieties, such as sialic acids, are often lost due to the high degree of vibrational excitation of the ions during ionization. To circumvent such issues, derivation techniques like permethylation have been developed. MALDI-TOF-MS is the most frequently used method for profiling permethylated glycans. The advantages of this method are that it enables the analysis of both neutral and acidic glycan structures to be performed simultaneously, the quantitation of glycans is reliable [222], and it is amenable to fragmentation analysis using tandem MS (e.g. TOF/TOF). Additionally, software (e.g. Glycoworkbench) and databases (Consortium for Functional Glycomics) have been developed to aid in the interpretation of MS spectra and MS^n fragmentation spectra. This tool facilitates rapid and accurate annotation of spectra and minimizes laborious fragmentation needed to characterize ambiguous peaks or perform *de novo* glycan structure assignment. Taken together, MALDI-MS is a valuable tool for large-scale diagnostics and biomarker studies and has been applied extensively to analyze the changes in the composition or abundance of N- and O- glycans present in the serum of cancer patients, or on the surface cancer cell lines[223].

Based on the strengths and advantages described above, a method using MALDI-TOF/TOF of analysis was developed. In this method, glycoproteins or proteins from the cell surface were reduced, denatured, and digested into peptide fragments. N-Glycans were released enzymatically using PNGaseF and O-linked glycans were released using Alkaline β-elimination. During processing N- and O- glycans are purified separately, and thus the N- and O-linked glycome are analyzed separately. Finally, the glycans were permethylated and then subjected to MALDI-TOF analysis. The resultant data was analyzed using a combination of expert curation, the CFG database, and GlycoWorkbench software[224,225]. In cases where it was not possible to identify the structure, fragmentation was performed.

Lectin array analysis

In addition to analytical tools using MALDI-TOF, Glycan binding proteins (GBPs) (such as lectins or anti-glycan antibodies) are valuable tools for glycomics analysis due to their specificity for defined glycan motifs and ability to capture structural information in the context of tissue. Lectins are non-immunological glycan binding proteins that exhibit specificity towards a particular glycan motif or structural feature [31,226]. For instance, the *Sambucus nigra* lectin (SNA-I) exhibits specificity for Neu5Ac α2-6-Gal(NAc), and is routinely used for the detection of terminal α2-6 linked sialic acid structures [227]. Currently, more than 100 lectins have been isolated and characterized. Compared to analytical methods that provide detailed information on glycan structure, lectin analyses provide an orthogonal measurement of the structural glycome based on broad structural features or motifs. One of the greatest strengths of lectins is their ability to measure certain structural features, such as monosaccharide linkage isomers or acidic glycans, much faster and easier than analytical techniques, which require special derivatization, purification, separation, and multiple rounds of MS [31,228].

Given the strength of these tools, a solid phase lectin array was developed to complement the structural data obtained using MALDI-TOF analysis. In the past, both well-plate and printed array formats have been used to characterize the glycome through motif profiling of cells and glycoproteins [227,229]. As lectin arrays do not require laborious glycan release or fractionation method for analysis, they can be performed in a high-throughput fashion. In the context of this study, a panel of biotinylated lectins were fixed to streptavidin coated well-plates and incubated with fluorescently labeled glycoproteins, or fluorescently cellular micellae preparations. These plates were washed and the fluorescent signal was quantified using a plate reader. Here the lectins are used as structural probes, whereby binding (or loss of binding) indicates the presence or absence of specific glycan motifs. The specific lectins selected, as well as their structural specificity, are noted in Table 7.1.

Transcriptomics analysis of glycan biosynthesis genes

Transcriptional analysis of the genes involved in glycan biosynthesis interactions provide another unique insight towards the characterization of the glycome. Integration of transcriptomics data with analytical structural analysis validates and complements the structural data obtained in the MALDI-TOF and lectin array analysis and enables inquiry into the complex biosynthetic regulation of glycans [30,230]. Additionally, identifying the glycogene changes responsible for changes to the glycome opens up new ways to probe the functional role of glycans at a cellular or organismal level through genetic manipulation. Animals and cell lines of glycosyltransferase and glycosidase knockouts provide another critical dimension towards understanding structure-function relationships.

To enable this method, a comprehensive list of N- and O-biosynthesis and degradation genes were compiled using literature search, the KEGG glycan database, and the Consortium of Functional Glycomics transferases interface. This list allows for the extraction of N and O-linked glycan biosynthesis genes from publically available transcriptomics datasets.

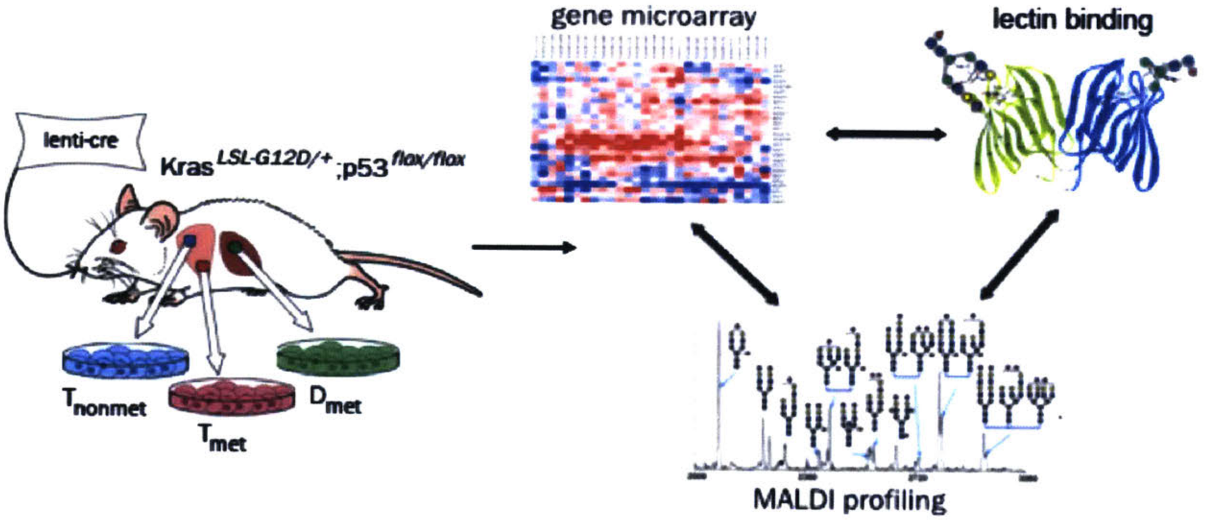


Figure 7.1 **General framework for the integrative glycomics analysis platform**. Cell lines derived from the Jacks Lab Kras$^{LSL-G12D/+}$; p53$^{flox/flox}$ mouse model of Lung Adenocarcinoma (hereafter KP model) were subjected to transcriptomic and glycomics profiling using MALDI-TOF. Lectin array analysis is used to cross-validate the glycomic data from MALDI TOF mass spectrometry. After identifying the glycomics changes across cell types, computational motif mining is done to identify glycan motifs that are differentially expressed.

Integrated approach to characterize the N- and O- linked glycans of a model cell system

The goal of this chapter is to implement an efficient and integrated approach to study the glycome of cells (Figure 7.1). Towards this, an integrated framework was implemented, including three assays:

1. N- and O- glycomics profiling using MALDI-TOF/TOF analysis of permethylated glycans
2. N- and O- glycan motif identification and quantitation using solid phase lectin array analysis
3. Transcriptomics analysis of glycogene expression.

Glycan isolation, Permethylation, MALDI-TOF/TOF, and lectin array methods were first developed and optimized using glycoprotein standards (for the sake of brevity, the data is not shown here).

To validate and investigate the utility of our integrated approach, a model cell line was selected. The work here was performed using cell lines derived from a tumor from a Kras $^{LSL\text{-}G12D/+}$; p53 $^{flox/flox}$ (hereafter KP) mouse model of NSCLC developed by Tyler Jacks [231]. These cell lines were derived from tumors representing the various stages of metastasis including, non-metastatic cells derived from a primary lung adenocarcinoma lesions (T_{nonmet}), metastatic cells derived from a primary lung adenocarcinoma lesion (T_{met}), and metastatic cells derived from a secondary metastatic site such as liver or draining lymph node (D_{met}). This model system was selected for several reasons: i) these cells models are derived from genetically identical tissue of origin, yet they possess different biological phenotypes (i.e. metastatic vs non-metastatic), ii) each cell model had a well characterized biological phenotype, and iii) transcriptomics data was readily available. Additionally, other research groups have performed 'omics level studies, such as ECM-binding array measurements [232]. Thus, by performing glycomics analysis and correlating it with other diverse datasets, it is possible to generate new hypotheses concerning structure-function relationship of glycans.

Experimental Methods

Cell Line Generation and Tissue Culture

The cell lines were derived from Kras$^{LSL-G12D/+}$; p53$^{flox/flox}$ (hereafter KP) mice described previously. Briefly, a lentiviral vector expressing Cre-recombinase was administered intratracheally to the KP mice. The mice develop lung tumors and develop macroscopic metastases in draining lymph nodes, pleura, kidneys, heart and liver. The KP model closely emulates the phenotype of human lung adenocarcinoma, and the selected mutational drivers (Oncogenic K-ras, and mutant p53) represent a significant fraction of the mutations observed in human lung adenocarcinoma (e.g. ~30% of NSCLC harbor mutations in the K-ras oncogene). Lung tumors and metastases were resected, digested and plated on tissue culture treated plastic. Because the lenti-viruses integrate stably into the genome, each tumor bears a unique molecular identifier marked by the lenti-integration site. Consequently, it is possible to unambiguously 'match' distant metastatic lesions to their primary lung tumor of origin. T_{nonmet} cells were isolated from primary tumors lesions that had not metastasized. T_{met} cells were metastatic cells isolated from the primary tumor site, D_{met} cells were isolated from a tumor lesion on distant sites, this cell line was clonally matched to the primary tumor metastatic cell line (T_{met}). These cell lines were cultured in Dulbecco's modified Eagles Medium (DMEM), 10% fetal bovine serum, penicillin/streptomycin, and glutamine.

Collection of N-and O- glycan biosynthesis genes

The list of N- and O-linked glycan biosynthesis genes were compiled from several sources including: The N- and O- biosynthesis/degradation list reported in [17,230], the KEGG glycan database(http://www.genome.jp/kegg/glycan/), and the Consortium of Functional Glycomics transferases interface (http://www.functionalglycomics.org/glycomics/molecule/jsp/glycoEnzyme/geMolecule.jsp). Some of the genes responsible for N and O- linked glycan biosynthesis were not represented in the microarray used to collect the transcriptomics data set and are excluded from our pathway analysis. The approved mouse gene symbols are listed in the Jackson Laboratory mouse gene name database (http://www.informatics.jax.org/mgihome/nomen/).

Gene Expression profiling and analysis

The expression of glycan related biosynthesis genes was analyzed using published microarray data from twenty-three primary tumor and metastasis derived cell lines. The RNA was isolated and hybridized to the Affymetrix GeneChip Mouse Exon 1.0ST array. Briefly, the image file was pre-processed using the aroma.affymetrix and FIRMA libraries available in the R/Bioconductor Software environment. The probe intensities were summarized as expression levels using quantile normalization and RMA. To identify the differentially expressed genes between T_{nonmet} and T_{met}/D_{met} samples, the significant analysis of microarray algorithm was applied. The gene expression microarray results that we mined can be found in [231]

N-linked and O-linked Glycan extraction, purification, and permethylation

$25x10^6$ cells were released from plate with TrypLE (Gibco, Cat no. 12605-036) and pelleted at 115xg for 10 minutes. The pellet was washed twice with PBS. The supernatant was removed and approximately 2mLs of extraction buffer (1%CHAPS, 5mM EDTA, 25mM Tris, pH 7.4) per $25x10^6$ cells was added to the cell pellet. The sample was sonicated (QSonica Q700 sonicator) on ice for 10 seconds (continuous mode, 40 Amps) followed by a 30 second break; this was repeated 5 times. The sample was transferred to Slide-A-Lyzer Dialysis Cassettes, 10K MWCO, presoaked in water. The sample was dialyzed in dialysis buffer (50mM Ammonium Hydrogen Carbonate) on a magnetic stirrer at 4°C cold room with 2-3 changes of buffer over a period of 48 hours. The dialyzed sample was transferred to a glass tube and lyophilized. The lyophilized sample was resuspended in 1.5mL of reduction buffer (2mg/mL DTT, 0.6M Tris, pH8.5) for 45 minutes at 37°C. Carboxymethylation was performed by the addition of 1.5mLs 12mg/mL IAA for 90 minutes at room temperature in the dark. The reaction was terminated by dialyzing in 50mM Ammonium carbonate buffer at 4°C for 16-24 hours; the sample was then lyophilized. The sample was resuspended in 1.3mg/mL TPCK bovine pancreas trypsin, 50mM Ammonium hydrogen carbonate, pH 8.4 solution such that the volume provides a 1:100 ratio of trypsin to protein (w/w). The sample was digested at 37°C for 4-8 hours. A few drops of 5% acetic acid was added to the sample and then purified using a Sep-Pak® C_{18} (Waters Corporation) column with the propan-1-ol / 5% acetic acid scheme (hereafter, P/A scheme). The P/A scheme includes: a) column conditioning through successive washes with 100% MeOH (5 mLs), 5% acetic acid (5mLs), propan-1-ol (5mLs) and 5% acetic acid (3x5mLs), b) direct loading of the sample to the Sep-Pak® C_{18} column, c) column washing with 5% acetic acid (30mLs) and d) stepwise elution with 3-4mLs each of 20% propan-1-ol/5%acetic acid, 40% propan-1-ol/5% acetic acid. The eluted fractions were collected, covered with pre-pierced parafilm and evaporated to dryness. The 20% propan-1-ol/5% acetic acid and 40% propan-1-ol/5% acetic acid fractions were combined in 200uL of 50mM ammonium hydrogen carbonate, pH 8.4. 3U of PNGaseF was added to the sample and incubated at 37°C for 20 hours. The digest was purified using a modified P/A scheme (omit column washing (c) and elute with 5% acetic acid (5mLs), 20% propan-1-ol / 5% acetic acid (4mLs)). The 5% acetic acid fraction contains the N-linked glycans, and the 20% propan-1-ol/ 5% acetic acid fraction contains the peptides/O-linked glycopeptides. Both fractions were covered with pre-pierced parafilm and evaporated to dryness. The O-linked glycans were released using reductive elimination. The sample was resuspended in 400 uL sodium borohydride solution (1M sodium borohydride, 0.05M NaOH) and incubate at 45°C for ~16 hours. The reaction was terminated by adding glacial acetic acid dropwise until fizzing stops. The sample was washed in a desalting column (see Preparation of Desalting Column in supplementary information section). Wash the column twice with 5% acetic acid, add the sample to the column, and then add 5% acetic acid; Collect ~ 5mLs of eluent; The sample was then lyophilized. As a note, never let the Dowex resin in the column run dry. To remove excess borates, the sample was resuspended in 10% acetic acid/ methanol, and evaporated to dryness under a stream of nitrogen at room temperature; this was repeated twice more. Both the N-linked and O-linked fractions were then permethylated. 0.5 mL of a DMSO/NaOH slurry (3 mL DMSO, 5 NaOH pellets- ground with a pestle) was added to the sample, followed by 0.2

mL methyl iodide. The mixture was mixed thoroughly and agitated on an automated shaker for 10 minutes at room temperature. The reaction was quenched by a slow, dropwise addition of ~1 mL of water with constant shaking between additions (to lessen the effects of the highly exothermic reactions). 1 mL of chloroform was added to the sample and water was added to bring the total volume up to 3mLs. The sample was mixed thoroughly, allowed to settle into two layers. The upper aqueous layer was discarded, and the lower chloroform was washed several times with water until completely clear. The chloroform layer was evaporated to dryness under a stream of nitrogen. The dried sample was purified using a Sep-Pak® C_{18} column using the aqueous acetonitrile system (hereafter, AA system). The AA system includes: a) Sep-Pak® C_{18} column conditioning by washing successively with methanol (5 mLs), Milli-Q water (5mLs), acetonitrile (5 mLs), and water (3x 5mLs), b) direct loading of the sample resuspended in 1:1 methanol water (200 uL), c) a stepwise elution with 5 mLs of water, and 2 mL each of 15%, 35%, 50%, and 75% aqueous acetonitrile. Each fraction was covered with pre-pierced parafilm and evaporated to dryness. The 35%, 50% and 75% contained the permethylated glycans and were pooled for MALDI-MS analysis. A detailed version of the general protocol can be found on the Consortium of Functional Glycomics (CFG) glycan profiling page (http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling-new.jsp).

MALDI-TOF Analysis of Permethylated N- and O- Linked Glycans

Permethylated N- and O-linked glycans were dissolved in 75 uL of methanol/water 8:2 (v/v) and mixed in a 1:1 ratio with 10mg/mL 2,5 dihydroxybenzoic acid in 80:20 (v/v) methanol/water. 1 uL aliquots were spotted onto a 384-well sample plate and dried under a desiccating vacuum. The analysis was performed using a Applied Biosystems 4800 Plus MALDI TOF/TOF Analyzer, equipped with a 355 nm Nd:YAG (neodymium-doped yttrium aluminium garnet) laser in positive reflector mode. The permethylated species were annotated with glycan structures according to the CFG Oligosaccharide Molecular Weight Search results. Assignments are based on compositions, taking into account biosynthetic constraints.

Lectin panel for lectin array analysis

Membrane vesicles from the lung adenocarcinoma cell lines were extracted and fluorescently labeled, as described previously [229]. A well-plate version of a solid-phase lectin array was developed. Briefly, a panel of biotinylated lectins was selected based on the motifs that they recognized (Table 7.1). Lectins were coated on a NeutrAvidin coated well-plates. Fluorescently labelled cellular micellae preparations were incubated in the wells, washed and imaged using a plate reader. The glycan binding motifs of each lectin were characterized using glycan array binding data available from the Consortium of Functional Glycomics (CFG).

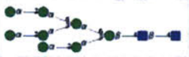
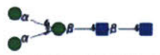
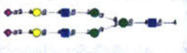
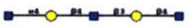
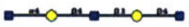
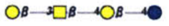
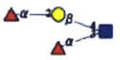
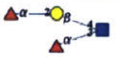
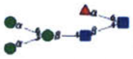
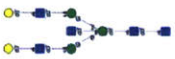
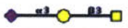
Lectin	Motif	Example
ConA	High-mannose	
HHL	High-mannose	
SNA	Terminal sialylation	
Jacalin	Long LacNAc chains	
GSL II	Long LacNAc chains	
PNA	GalNAc	
AAL	Fucose	
LTL	Fucose	
LCA	Core Fucose	
PHA-E	Bisecting GlcNAc	
MAL-II	O-glycans	

Table 7.1 **Lectin panel used in the lectin array analysis.** Additionally, the characterized motif along with an example glycan structure containing that motif are shown.

Results

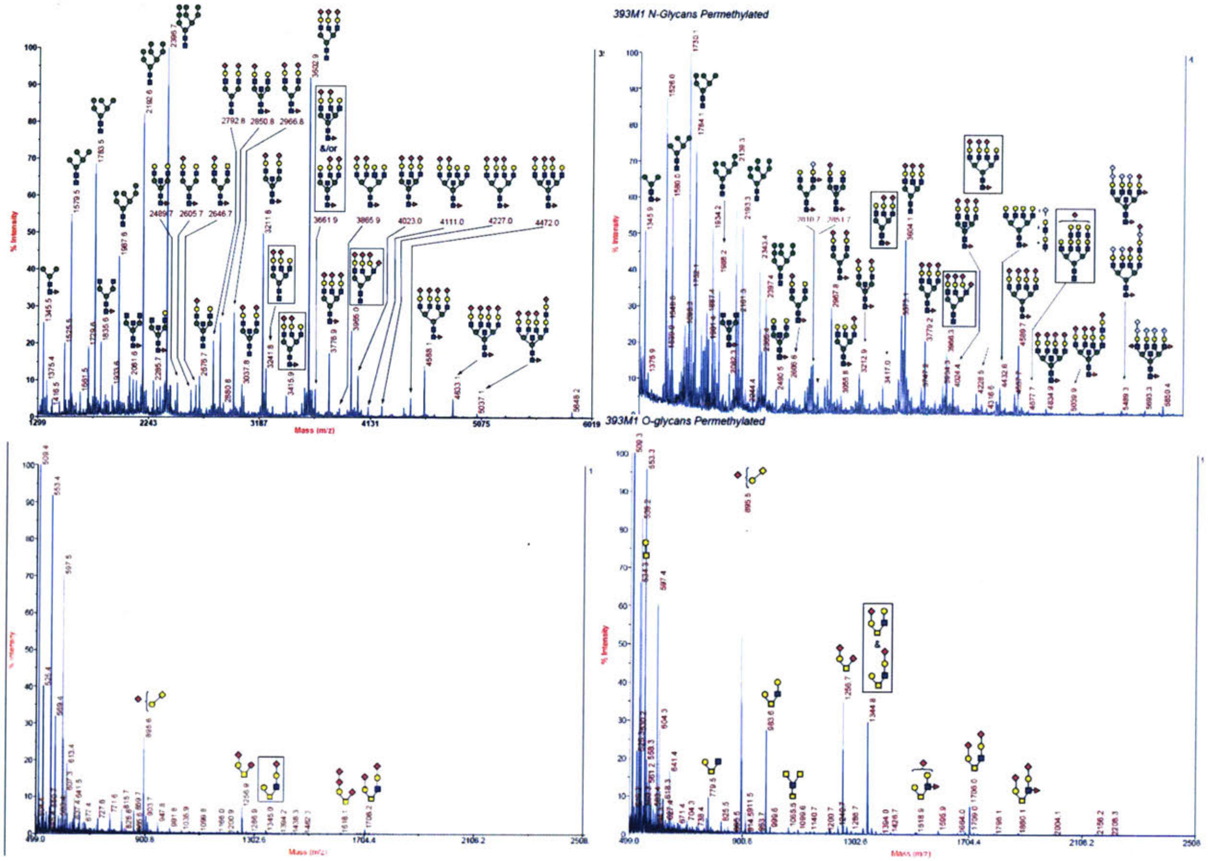


Figure 7.2 **Representative annotated MALDI-TOF spectra.** N-glycan spectra are shown in the top two panels, and O-glycan spectra are in the bottom two panels. The N and O- linked glycan spectra in left column are from Tnonmet cells, and the right column are from Dmet cells. The spectra are represented as relative intensity (with 100% being the most abundant structure in the pooled analysis) vs mass (m/z). Each peak represents a glycan structure, and the annotated structures are presented as cartoon (using CFG nomenclature).

To assess glycomic changes in these cell lines, glycans were extracted from Tnonmet, Tmet or Dmet cells. The glycans were isolated and purified using a combination of enzymatic and chemical methods, and were subsequently permethylated. The final samples were run on a MALDI TOF/TOF instrument and the m/z peaks were annotated (Figure 7.2). Importantly, some consider MALDI-TOF analysis of permethylated glycans to yield semi-quantitative data, and therefore the peak heights between spectra can be directly compared. In our view, a more robust method is to consider the MALDI-MS spectra to be a snapshot of the total glycan pool. In this case, the sum of all glycan peak heights equals the total pool and the abundance of each glycan can be expressed as the percent abundance in the glycan pool. In this way, shifts in relative abundance of glycans in the context of the total glycan milieu can be compared between two samples.

As can be seen in Figure 7.2 there were appreciable differences in N- and O-glycans between the two glycan spectra from T_{nonmet} vs D_{met} cells. To gain insight into how the glycan milieu changed, specific glycan structural features were selected and compared between non-metastatic cells (T_{nonmet}), and metastatic cells (T_{met} & D_{met}). The features selected for analysis were: bisecting GlcNAc, terminal sialylation, number of branches, branch length (number of LacNAc repeats), and core fucosylation. To capture the relative abundance of these features, the percent abundance of each feature was calculated by summing up the peak intensity of all the glycans containing a specific motif. These were expressed as a percent abundance of the total glycan pool. Thus, the relative abundance of each glycan motif could be compared in non-metastatic cells vs metastatic cells. The results of this analysis are summarized in Table 7.2 under the "MALDI profiling column".

Based on the MALDI profiling, the N-linked glycome of the metastatic cells showed a shift in the high mannose glycans, namely, an increase in abundance of smaller high mannose glycans (<5 mannose residues). Additionally, an increase in branch length, core fucosylation, chain fucosylation was observed. Conversely, metastatic cells appear to have a decrease in terminal sialylation. Furthermore, metastatic cells showed a striking increase in the abundance of truncated O-glycans, and there was the expression of core 1 and core 2 glycans.

To corroborate the changes in the abundance of these glycan motifs, these cell lines were subjected to a lectin array analyses (Figure 7.3). For this analysis, fluorescently labeling micelles were created and applied to a well-plate coated with various lectins. A panel of lectins capable of detecting these previously observed glycan motifs were selected (Table 7.1). Generally, the lectin array results are in agreement with the MALDI analysis. AAL and LCA, recognizing fucose and core fucose respectively, showed significantly increased binding to metastatic cells (Figure 7.3). Given the correlation with the MALDI data, this builds confidence in our conclusion that metastatic cells have increased abundance of core fucosylation and chain fucosylation.

Additionally, Jacalin and GSL-II are known to recognize long LacNAc chains, and showed significantly increased binding to metastatic cells, further supporting the notion that metastatic cells express glycans with increase branch length.
Furthermore, PNA, which binds to T antigen (a core 1 O-glycan, Gal-β(1-3)-GalNAc), showed a small increase in binding to metastatic cells.

Interestingly, some features were not well-correlated with our MALDI-guided glycomics feature analysis. For instance, no significant changes were detected in high mannose structures or terminal sialylation. As an aside, a lectin-based flow cytometry analysis of intact cells was also developed (not discussed here). Using this method, SNA-I (a sialic acid binding lectin) showed decreased binding to metastatic cells. It is possible during the process of generating the labelled cellular micelle extractions, changes occur to the sialylated glycans which limited our ability to capture changes in the solid phase lectin array format.

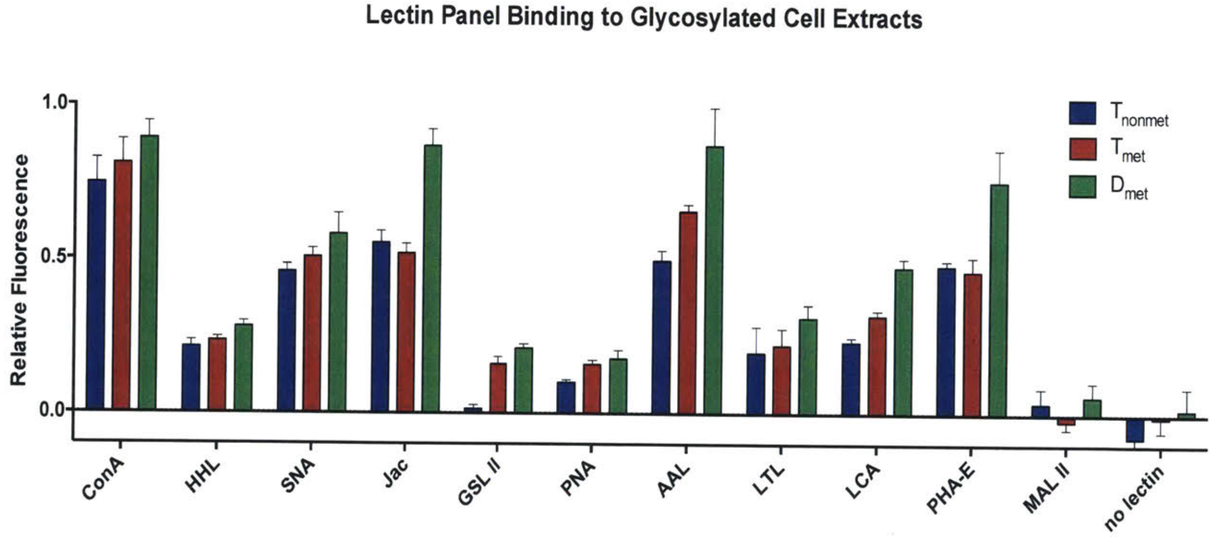


Figure 7.3 **Lectin array results**. The average of three replicates are shown here; Error bars = standard error of the mean. Tnonmet cells are represented as blue bars, Tmet cells are represented as red bars, and Dmet cells are represented as green bars.

Next, focused analysis of N- and O-linked glycan biosynthesis genes was performed. Using publically available databases, glycan biosynthesis gene networks were mapped out and a comprehensive list of N- and O-linked glycan biosynthesis genes was assembled. Gene expression data from T $_{nonmet}$, T $_{met}$ & D $_{met}$ was published previously. A comprehensive assessment was performed (Figure 7.4, Figure 7.5). In order to identify the gene expression changes that underlie the differential abundance of glycomic structural features, the genes involved in the synthesis of these features were analyzed. In certain instances, the relationship between the glycans expressed and the underlying biosynthetic was straightforward. For instance, core fucosylation is increased in metastatic cells. This feature is catalyzed by Alpha1,6-focosyltransferase (*Fut8).* As expected, analysis of glycogene expression revealed that expression of *Fut8* is increased in metastatic cells. In other instances, glycan features emerge due to a complex interplay of biosynthetic genes. An increase in abundance of Poly-LacNAc extensions could result from many enzymes including those that catalyze the GlcNAc from the N-glycan core structure (MGAT4a), enzymes that catalyze LacNAc extension (i.e. B4GALT1&4, B3GNTs) and those that terminate LacNAc extension by capping with sialic acid (i.e. ST6GALNAcs and ST3GALNACs) [2]. Importantly, there is competition between GlcNAc extension enzymes and sialyltransferases, so the ratio in expression is better correlated with the observed features. In this model system, metastatic cells showed an increase in *Mgat4a* gene expression and concomitant decrease in *St3gal1 & 4, Gcnt3* involved in GlcNAc capping. Finally, there are some cases where no correlation was observed between biosynthetic genes, and the resultant glycan structure. For instance, chain fucosylation is regulated by FUT1, FUT2, and SEC1 which catalyze the addition of a fucose residue to N glycan chains. Both MALDI-MS analysis and the lectin array data suggest that there is an increase in chain fucosylation yet in metastatic cells show a decrease in *Fut2* and *Sec1* expression. Other integrated transcriptomics/glycomics studies reporting similar results and attribute this to regulatory mechanisms that can't be captured using transcriptomics alone [17,230]. The remainder of the glycogene transcription analysis is detailed in Table 7.2

Glycan Feature	MALDI Profiling	Lectin Binding	Glyco-Gene Analysis	Conclusion
High mannose		n.s.	complex relationship	Shift in distribution of high mannose structures
≤5 Mannose	↑			
>5 Mannose	↓			
Long / high order branches	↑	↑	*Mgat4a* (↑) *St3gal1 & 4, Gcnt3* (↓)	Overall increase in number of LacNAc repeats
Terminal sialylation	↓	n.s.	*St3gal1 & 4* (↓)	Decrease in terminal sialylation
Chain fucosylation	↑	↑	*Fut2, Sec1* (↓)	Overall increase in fucosylation, despite decrease in *Fut2, Sec1*
Core fucosylation	↑	↑	*Fut8* (↑)	Increase in core fucosylation
Core 1 & Core 2 O-glycans	↑↑	↑	*Galnt1,7,9,13, Gcnt1* (↑) *Galnt4, Gcnt3* (↓)	Increase in core 1 creation, *Gcnt1* appears to dominate core 2 creation

Table 7.2 **Summary table of the integrated glycomics analysis.** The data is presented relative to the non-metastatic cell line, where ↑ indicates that a feature is more abundant in the glycan pool of metastatic cells (↓ denotes the opposite); n.s.= not significant

Conclusion

Here, an integrated approach to study the glycome (Figure 7.1) was developed. To investigate the utility of this approach (and further validate our methods), these tools were applied to a non-metastatic vs metastatic cell model in an effort to characterize the structural changes that occur to N- and O-linked glycans during metastatic progression. The approach developed here can identify changes to the glycome at multiple levels (i.e. changes in abundance of specific glycan or changes to glycan motifs), and link these changes to the glycan biosynthesis genes responsible. Thus, these methods provide a robust and efficient way to characterize the glycome, and provide a key first step towards decoding structure-function relationships of glycans.

Furthermore, although it was not the explicit focus of this study, several interesting biological observations were made. Metastatic cells showed an increase in core fucosylation. Interestingly, increases in expression of *FUT8* and core fucosylation of N-glycans has been reported in lung and breast cancer in humans [233]. Furthermore, core fucosylation of alpha-fetoprotein is a specific biomarker of carcinomas and is approved for the early detection of hepatocellular carcinoma [234] . Increases in core fucosylation are thought to drive tumor progression through altering cell-ECM interaction (i.e. regulation of migration via $\alpha_3\beta_1$ integrin function[235,236]), and inducing changes in cell signaling (i.e. increases in core fucosylation increases dimerization and activation of EGFR [237]). Additionally, several metastasis-associated O-glycan changes were identified, namely increase in the expression of core 1 and core 2 O-glycan structure. Several key tumor associated carbohydrate antigens were increased, including T-antigen (Gal-β(1-3)-GalNAc) and di-sialylated T antigen. These truncated O-glycans are known to be specific to cancer cells, and are observed on almost all epithelial cancer cells [238,239]. These truncated core 1 and core 2 structures are known to play critical roles in proliferation, differentiation, and migration. Additionally, increases in expression of these structures on normal cells are known to directly induce oncogenic features [238]. Taken together, it appears that the KP model of NSCLC recapitulates several key glycomics features of human cancer. Thus, this system could be useful to glycan-mediated cell-microenvironment interactions, or develop anti-metastatic therapies targeting glycan mediated interactions.

It is worth noting that most cancer glycomics studies use human sample or human tumor cells lines, focusing on a rather passive cataloging of the glycomics changes versus normal tissue. Very few cancer glycomics studies have been performed using genetic animal models of multistage cancer progression and spontaneous metastasis. Given the critical role that glycans play at the cell-ECM interface, there is a great need to dissect the structure-function relationships of glycan that mediate metastatic progression. This animal model and the information obtained through this integrated glycomics analysis allow for an unprecedented opportunity to intimately dissect the glycan mediated cell-microenvironment interactions critical for metastatic progression. The development of this integrated glycomics analysis framework is a critical first step towards this goal.

In this vein, other research groups had performed ‘omics level studies using this same cell lines, the characterization of the N- and O-linked glycome now enables the ability to generate novel structure-function hypotheses about the biological role of glycans in metastasis. Through correlation of diverse systems-level data sets, it may be possible to yield new biological insights into the biological function of glycans. For example, previous studies used an ECM-binding array to capture the changes that occur in cell-ECM interaction during metastasis. In this study Tnonmet, Tmet and Dmet cell lines were tested for their adhesion to ECM on an ECM array platform [232]. This array revealed that galectin-3 or -8 conferred stronger adhesion in metastatic cell lines. Interestingly, Galectins -3 and -8 are glycan-binding proteins which recognize long LacNAc repeats (galactose-N-acetylglucosamine)[232]. From our integrated analysis, metastatic cells were found to have an increased abundance of N-linked glycans with elongated poly-LacNAc repeats. It is possible that the glycomics changes drive the increase in adhesion to Galectin-3 and -8. If this were true, this finding could have important therapeutic implications. Further experimentation would be required to test this hypothesis, but this exemplifies how integrating functional assays with glycomics analysis yields new insights into structure-function relationship of glycans.

In the broader context of this thesis, this study highlights the need for and the utility of an integrated approach to study glycomics. Furthermore, although this section focused on developing an approach to study N- and O-linked glycans, the framework, expertise and insights developed here are generalizable to the study of other glycans.

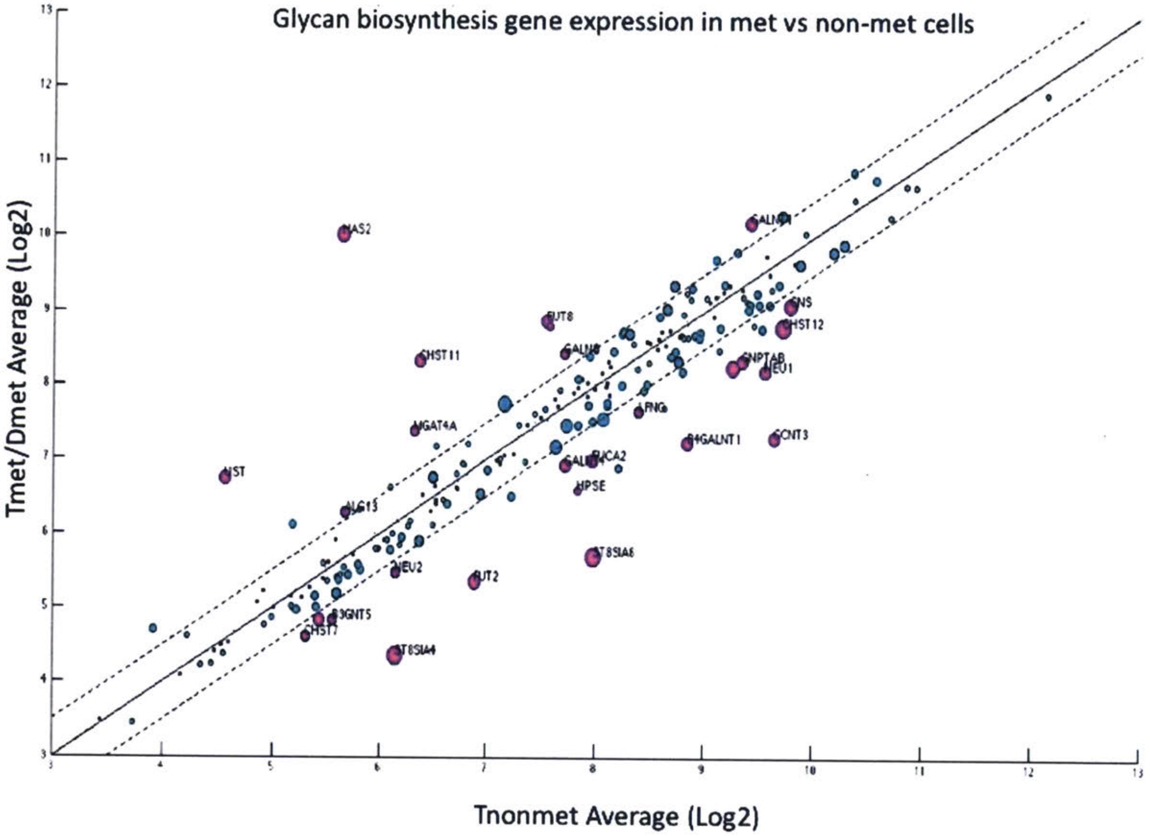


Figure 7.4. **Glycogene expression bubble plot in Metastatic cell vs Non-metastatic cell lines from KP mice.** Each bubble represents one glycan biosynthesis gene. Not that this analysis here included N-linked, O-linked and GAG biosynthesis genes. For some genes, the gene name is listed next to the bubble. Pink bubbles are genes that had significantly different expression (measured by the Significant Analysis of Microarray statistical technique). Additionally, size of the bubble plot scales inversely with the log of the p value. Glycogenes above the solid black line are expressed higher in Metastatic lines, and those below the solid black line are expressed more highly in non-metastatic cells. The dotted line demarcates a cutoff of + or - 0.5 log2 expression from the solid black line.

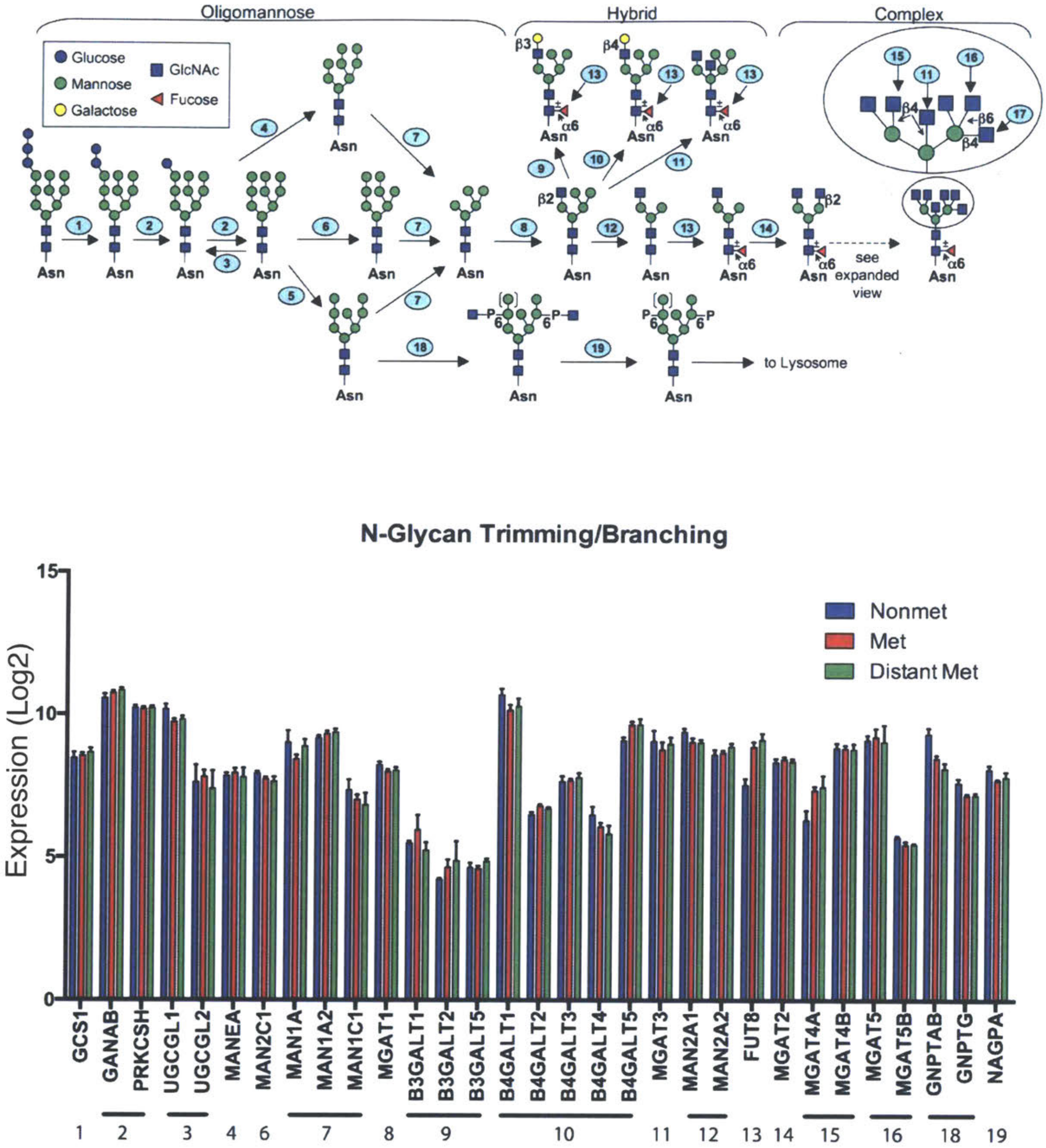


Figure 7.5 **A representative example of the integrated biosynthetic gene network and gene expression data**. Shown here on the top panel is the biosynthetic network regulating N-glycan trimming and branching. The number in the blue circle correspond to one or more enzymes responsible for this reaction. The numbers in blue circles correspond to the numbers below the bar graph. For the bar graph, data is expressed as the log 2 average of gene expression Non-metastatic, metastatic and distant metastatic cell lines (N= 4 lines, N=7 lines, N=3 lines respectively). Error bars represent standard error of the mean.

This work resulted in the following publication:

"Functional Glycomics Analysis: Challenges and Methodologies" published February 2, 2014 in Fundamentals of Advanced Omics: From Genes to Metabolites, Volume 63, Pages 255;

Abstract

Glycan are a ubiquitous class of biological molecules responsible for modulating a diverse array of physiological processes including development, immune recognition, and host–pathogen interaction. Elucidating structure–function relationships has been challenging due to the structural complexity of glycans and the dominant role of multivalency in their biological interactions. Therefore, a glycomics approach, defined as an integrated, systems-level study of glycans, is necessary to elucidate the structure–function relationships. In this chapter, we discuss the fundamental glycomics tools used for structural elucidation and functional analysis across multiple levels, emphasizing integration across these axes. We conclude with a case study of influenza A virus hemagglutinin interaction with sialylated glycan receptors to demonstrate a robust structure–function paradigm that is born out of a glycomics approach.

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Devin Quinlan, Rahul Raman, Jing Li, Prithwis Mukhopadyay and Ram Sasisekharan for their contributions to this work. This work was funded in part by Stand Up to Cancer (SU2C/AACR), the Koch Institute CTC Project, National Institutes of Health (T32 ES007020-38) and MIT UROP Funding Office

Chapter 8 : SULF1 modulates the activity of breast cancer stem cells

Summary and Significance

Breast cancer is the second leading cause of cancer-related deaths among women in the US. Almost all cancer related deaths are due to metastasis and resistance to therapy. The latter has been attributed to the existence of cancer stem cells (CSCs), a cell population characterized by their increased resistance to chemotherapeutics and ability to seed new tumors and metastasis. Therapeutically targeting CSCs in tumors could lead to a durable clinical response and prevent metastatic disease. In this context, there is great need to uncover the CSC-specific molecular pathways that regulate CSC maintenance and activity.

Interactions between CSCs and their microenvironment are important for CSC behaviors such as metastasis, chemotherapeutic resistance and tumor initiation. Despite knowledge that HSGAGs play a critical role in cell-microenvironment interactions, the function of HSGAGs in cancer stem cell activity is largely unknown. Recently, analysis of microarray data showed that an HSGAG modifying enzyme, SULF1, was highly expressed in breast stem cells (Unpublished Observation, Weinberg lab). SULF1, which is an extracellular 6-O-endosulfatase, acts as regulator of cell-microenvironment signaling, and is known to directly interact with and regulate several of the known autocrine and paracrine signals involved in breast stem cell and breast cancer stem cell activity. Thus, we sought to conduct the first investigation into how a HS-modifying enzyme, SULF1, affects breast CSC behavior through modulation of cell-surface HSGAG sulfation.

Leveraging the framework developed in Chapter 7, we apply an integrated glycomics approach to uncover the structure-function relationships of HSGAGs in CSC activity. Here, we report that SULF1 is required for tumor initiation, growth and metastasis of CSCs. Furthermore, we find that SULF1 plays a critical role in maintaining the EMT'd state of CSCs and that knockdown of SULF1 expression induces a mesenchymal to epithelial transition via increasing 6-O-Sulfated HSGAGs which negatively regulate WNT and TGFβ signaling.

Finally, we report that SULF1 is highly upregulated in tumors from patients with invasive breast carcinoma. Moreover, patients with ER-/PR- Basal breast cancer that express SULF1 highly have a poorer relapse free survival. These results suggest that SULF1 might play a role in human tumor progression. Taken together, the identification of SULF1 and its essential role in CSC physiology and regulation of the EMT suggests that targeting this enzyme might be a viable therapeutic strategy to reduce CSC activity.

Introduction

Breast cancer is the second leading cause of cancer-related deaths among women in the US [240]. Triple negative breast cancers (TNBC) (defined by the absence of HER2 amplification, the Estrogen receptor (ER) and Progesterone receptor (PR)) accounts for ~15% of all breast cancers [241]. TNBC is a highly aggressive subtype and is clinically challenging to treat. Currently, the standard of care treatment for TNBC is a chemotherapeutic agent, however, the response is poor and patients often relapse or succumb to metastatic disease [67]. Many hypothesize that the inability to treat TNBC is due to a high degree of intra-tumoral heterogeneity, specifically, our inability to kill the aggressive cancer stem cell (CSC) fraction of the tumor. CSCs are functionally defined by their ability to seed new tumors, including metastases, and are often more resistant to chemotherapy than the bulk of neoplastic cells within a tumor. It is possible that therapeutically targeting CSCs in tumors would lead to a durable clinical response and prevent metastatic disease. In this context, there is great need to uncover the CSC-specific molecular pathways that regulate CSC maintenance and activity.

Interactions between CSCs and their microenvironment are important for CSC behaviors such as invasion, migration, resistance to anoikis, chemotherapeutic resistance and tumor initiation [68]. The CSC niche is highly complex and is composed of numerous molecular players/interactions including: Hormonal signals, paracrine and autocrine signals such as morphogens/cytokine/growth factors, extracellular matrix molecules that directly interact with CSC [69–71]. Additionally, cellular players such as activated stroma, immune cells, or surrounding tumor cells create/alter the microenvironment and ultimately direct CSC activity [64]. Preclinical evidence suggests that targeting CSC-microenvironment interactions in cancer might reduce the CSC number or activity thereby improving disease outcomes[72]. It remains unknown whether or not targeting CSC-microenvironment will yield clinical benefit and numerous clinical trials are underway testing this hypothesis [73]. Thus, the microenvironment of CSCs represents fertile ground for the discovery of new therapeutic targets.

There is growing evidence demonstrating that cell-microenvironment interactions, and ultimately cellular function/phenotype, are strongly influenced by heparan sulfate glycosaminoglycans (HSGAGs or HS) present on the cell surface and in the ECM surrounding cells. HSGAGs are complex, linear polysaccharides comprised of repeating disaccharide units of uronic acid linked to glucosamine residues[4]. They are attached to a proteoglycan core and are responsible for mediating many critical processes at the cell-ECM interface and within the ECM itself. During HS biosynthesis, polysaccharide chains can be differentially sulfated or acetylated, giving rise to diverse patterns of modifications on the HSGAGs [85]. HS chemical diversity enables their interaction with a diverse array of growth factors, cell-surface receptors, and ECM components, thereby modulating signaling and, in turn, cellular phenotype. Thus, HSGAG chains on heparan sulfate proteoglycans (HSPG) can (i) act as co-receptors for growth factors/chemokines to facilitate oligomerization or receptor-ligand complex formation, (ii) create morphogen or growth factor gradients through affinity-based localization of HS binding components, and (iii) store or sequester growth factors/enzymes in the ECM[86].

The regulation of HSGAG modification is highly dynamic and can be altered depending on cellular or disease state. Sulfation of HSGAGs is modulated biosynthetically by sulfotransferases (which add sulfate groups) and post-synthetically by endosulfatases (which remove sulfate groups)[85]. Recently, two extracellular endosulfatases, SULF1 and SULF2, have emerged as key regulators of growth factor signaling, demonstrating critical roles in development, tumor growth and metastasis[5,88,94]. SULFs remove the 6-O sulfate (6-O-S) modification of the glucosamine with a preference for the tri-sulfated disaccharides[242]. The expression of SULFs can positively or negatively regulate signaling depending on the nature of the growth factor-HSGAG interaction[95,96]. For instance, in the case of FGF-2 signaling, 6-O-S modified HS is required for the formation of the active FGF-2:HSGAG:FGFR1 signaling complex [97]. In multiple myeloma, the experimentally induced expression of SULF1 reduced tumor growth *in vivo* and reduced the formation of FGF:FGFR1 ternary complexes, implicating SULFs as suppressors of tumor formation or growth[7].

Acting in the opposite direction are WNT ligands, which demonstrate high-affinity binding to 6-O sulfated HS; this binding prevents functional interaction of the WNTs with their cognate cell-surface Frizzled receptors[89]. In this case, high expression of SULFs would reduce the overall 6-O-S modification of HSGAGs, thereby liberating previously sequestered WNTs and enabling binding to the receptor Frizzled. In this way, SULFs are thought to act as oncogenes, mediated through alterations in autocrine WNT signaling [98].

The 6-O modification of HS has also been shown to influence a variety of other growth factor ligand/receptor interactions, including those involving the BMP, HGF, HB-EGF, and VEGF ligands[95]. In human tumors, SULF expression is highly variable and has been shown to behave either as a tumor suppressor or as an active promoter of tumorigenesis and metastasis. Given this functional ambiguity, the latter observation highlights the importance of considering the micro-environmental context when dissecting the function of HSGAGs.

Recently, several reports characterizing the intracellular, autocrine, and paracrine signal that regulate stem cell/CSC maintenance and activity have been published [69,74]. However, there have been few studies on how HSGAGs, and the enzymes that modulate their fine structure might play a pivotal role in CSC activity. Recently, analysis of microarray data comparing breast stem cells vs non-stem cells revealed that SULF1 was highly expressed in breast stem cells (Unpublished Observation, Weinberg lab). Given what is known about SULF1 as an extracellular pleiotropic regulator of cell-microenvironmental signaling we sought to conduct the first investigation into how a HS-modifying enzyme, namely SULF1, affects breast CSC behavior through modulation of cell-surface HSGAG sulfation.

Experimental methods

Capillary Electrophoresis of Heparan Sulfate Dissacharides

Cells ($1x10^8$) were harvested from plates and incubated with 2.5X trypsin (Gibco Cat no. 15090046) at 37°C for ~1hour. Cells were pelleted at 4000xg for 30 minutes, and the supernatant was isolated for further analysis. The supernatant was filtered using a 0.45um syringe filter and spiked with Benzonase (Sigma cat no. E1014-5KU) at a concentration of 1:100 and incubated overnight at 37°C. Proteins were digested using Proteinase K (~ 1mg/mL) at 55°C for 2-3 hours. Note $CaCl_2$ was spiked into the sample to a final concentration of 5mM for optimal activity of Proteinase K. HSGAGs were isolated using anion exchange spin columns (Vivapure Q MAXI H Sartious cat no. VS-IX20QH08). The spin columns were washed in 20mM Tris pH 7.4, and eluted with 20mM Tris pH 7.4, 1M NaCl. The samples were buffer exchanged and concentrated into HPLC grade water using a 3000 MWCO filter. Exhaustive depolymerization of HS was performed using ~400 mU of Heparinases I, II and III incubated overnight at 30°C. Heparinases were removed using a Nickel spin column and the sample was concentrated. The disaccharides were analyzed using capillary electrophoresis using a high sensitivity flow cell in reverse polarity as described previously [107].

Animal Studies

All mouse studies were performed under the supervision of MIT's Division of Comparative Medicine in accordance with protocols approved by the Institutional Animal Care and Use Committee. NOD/SCID mice were bred in house. Mice were 2–4 months of age at time of injections. Tumor cells were resuspended in 20% Matrigel/MEGM (20 μl) for mammary fat pad injections. Tumors were dissected at the end of the experiment and weighed. GFP-positive lung metastases were counted from individual lobes by fluorescent microscopy.

Extreme limiting dilution analysis

The data generated from the limiting dilution assay was analyzed using the Extreme limiting dilution analysis software tool (http://bioinf.wehi.edu.au/software/elda/) [243]. Data are presented as cancer stem cell frequency with the 95% confidence interval.

Tumor Digests

Tumors were chopped into small pieces in sterile conditions then incubated at 37°C for 4 hours in DME containing Collagenase A and Hyaluronidase. Following digestion, tumor cell suspensions were pelleted, the DME removed, and then resuspended in 0.15% trypsin for 3 minutes. Trypsin was quenched with 10% IFS in DME. Cells were spun down then analyzed by FACS or frozen in freezing media (10%DMSO, 90% Calf Serum).

Cell lines and cell culture methods

The work in this text uses primary human mammary epithelia cells (HMECs), that have been immortalized with human telomerase (hTERT). Here, these cells lines are termed HMEs [244,245]. Recently, a floating cell population (HME-flopc) was isolated and characterized. This floating population was shown to be enriched for CD44^{hi}CD24^{lo}ESA- cells which are thought to be bipotent progenitors/stem cells. HME-flopc- CD44^{hi} possessed properties of stem-like cells and were capable of reconstituting ductal structures in a humanized mouse mammary fat pad model. Single cell clones of these HME-flopc- CD44^{hi} were generated, transformed using SV40 largeT and H-RAS- this derivative is called HMLER-flopc-CD44^{hi}. Importantly, in the context of the assays performed here the HMLER-flopc-CD44^{hi} clones were functionally indistinguishable from cells derived from HMECs that were transformed with SV40 largeT and H-RAS (termed HMLERs) where the CD44^{hi} population was purified using FACS (termed HMLER-CD44^{hi}). ShRNA knockdown of SULF1 was performed in a single cell clone of HMLER-flopc-CD44^{hi}.

HMLER CD44^{hi}, CD44^{lo} cells, and SULF1 shRNA knockdown cells were cultured using a modified MEGM media- 500mLs MEGM, 250mL F12, 250mL DME, and supplemented with B12, Insulin, hydrocortisone, EGF, G418, and pen/strep. Note that shSULF1 hairpin was doxycycline inducible, therefore doxycycline was added to the media. BPLER CD44^{hi} and CD44^{lo} cells were cultured in WIT media.

Mammosphere Culture & growth assays

Mammosphere cultures were performed as described in [246]. Briefly, cells (100) were seeded per well in a 96-well Corning Ultra-Low attachment plate (Corning, USA; CLS3474) in replicates of 10-16 wells; sphere numbers were counted between days 8 to 12. Cell growth assays were measured using cell-titre glo (promega).

RNA Isolation and qRT-PCR Analysis

Total RNA was isolated using the RNeasy Micro kit (QIAGEN). Reverse transcription was performed with miScript II RT Kit; miScript and Qantitect Primer Assays were used to detect miRNAs and mRNA (QIAGEN).

Vectors, lentivirus production and Viral infections

pBabe SV40-ER (Zeocin), pBabe H-Ras-GFP (Puromycin), production of virus and infection of target cells have been previously described in [244,247]. Infected cells were selected with Zeocin (100 µg/ml) and Puromycin (2 µg/ml). Dox-inducible shRNAs targeting SULF1 were purchased from Open Biosystems.

Flow Cytometry and FACS

Cells were prepared according to standard protocols and suspended in phosphate-buffered saline (PBS). For detection of CD44, cells were incubated with CD44-APC or CD44-PE-Cy7 (Biolegend, USA) at a 1:400 dilution for 30 minutes on ice, then washed in cold PBS and measured. For detection of heparan sulfate, cells were incubated for 45 minutes on ice with a 1:50 dilution of mouse 10E4 antibody (Seikagaku

America). After washing cells were stained with an APC-conjugated anti-mouse IgM Immunofluorescence for 30mins on ice. Cells were then washed three times in PBS and analyzed (or sorted). Cells were sorted on BD FACSAria SORP and analyzed on BD LSRII, using BD FACSDiva Software (BD Biosciences, USA).

Immunofluorescence

Tumor samples were formalin fixed, paraffin embedded, and cut into 5-µm sections. Antigen retrieval was performed with citrate buffer (pH 6.0) followed by boiling. Primary antibodies were used to detected E-cadherin (Cell Signaling) and SV40-LargeT (Santa Cruz Biotechnology). Alexafluor-594 and -488 secondary antibodies (Invitrogen) were used for detection.

Statistical analysis

Data are presented as the mean ± SEM. Student's t test was used to compare groups, $p<0.05$ was considered significant unless otherwise noted.

Results

Initial observations & Cell model system

Earlier in the field of cancer biology, there was much confusion regarding the origins of a cancer stem cell- was a cancer stem cell an adult stem cell that acquired mutations that enable malignancy or did cancer stem cells arise from a more differentiate epithelial cell which acquired properties of a stem like cell [248]? The answer to these questions have important implications for how one considers creating appropriate cell model system of cancer stem cells. In addressing this question, in the context of breast cancer, new evidence suggests that there are populations of normal, non-stem cells, that have the ability to spontaneously convert into stem-like cells; this holds true for neoplastic cells as well [119]. Furthermore, it was demonstrated that the naturally occurring stem cells were functionally equivalent to stem-cells that arose through conversion of non-stem to stem cells. This held true with cancer stem cells as well, where CSCs that spontaneously arose from transformed non-stem mammary epithelial cells were functionally equivalent to CSCs generated by the transformation of mammary stem-like cells [119]. These results have several implications for the study of cancer stem cell biology: i) These results suggest that transformation of stem-cells represent a good model of cancer stem cells, ii) additionally, this suggests that the study of signals that regulate stem-cell activity may also play a role in CSC activity. The latter has been known for some time as surface markers used to isolate stem-cells can also be used to isolate CSCs.

Towards understanding the molecular signaling networks that underlie the activity of cancer stem-cells, gene expression analysis was performed comparing normal stem like cells (HME-flopc-CD44^{hi}) to their parental line HMEs (comprise of >99% non-stem cells). The second most upregulated transcript in HME-flopc-CD44^{hi} cells was SULF1, a heparan sulfate 6-O-endosulfatase, which modified cell surface HSGAGs (unpublished observation, Weinberg lab). It is possible that SULF1 might be a good marker of stem-cells or cancer stem cells. Further supporting this hypothesis, analysis of tissue array data from the human proteome atlas shows that SULF1 expression in normal human breast tissue is restricted to the myoepithelial cells, where other stem cell markers (CD44) are expressed[249].

As discussed in the introduction, cancer stem cell maintenance and activity depends on autocrine and paracrine signals from the microenvironment [69,74]. Given the critical functional role that HSGAGs play in regulating cell-microenvironment signaling, we hypothesized that SULF1 might play an important role in modulating the activity of cancer stem cells.

SULF1 is a functional CSC Marker

To investigate if SULF1 was a functional marker of CSC, we used a well-established cancer cell model system- HMLERs (see methods for description). In HMLERs, cancer stem cells can be isolated, using FACs, based on their high expression of CD44 (where CSCs are denoted HMLER-CD44^{hi})[250]. HMLER-CD44^{hi} are functionally defined as cancer stem cells based on two key properties: i) their ability to form tumors efficiently when implanted into NOD/SCID mice at limiting dilution, and ii) their *in vitro* ability to resist anoikis and undergo anchorage-independent proliferation when seeded at clonal densities. Furthermore, when implanted into the mammary fatpad of NOD/SCID mice HMLER-CD44^{hi} cells form aggressive tumors which spontaneously metastasize to the lung. Using qRT-PCR, we confirmed that SULF1 was highly expressed in HMLER-CD44^{hi}, relative to the HMLER-CD44^{lo} (Figure 8.1a). Next, we wanted to further investigate the expression of SULF1 in variety of other cell models. Importantly, all of the cell lines selected represent model systems for triple negative breast cancer (TNBC) [251]. We investigated (BPLERs; BPLER CD44^{hi} vs CD44^{lo}), and found that SULF1 was highly expressed in the BPLER-CD44^{hi} CSC fraction relative to CD44^{lo} (Figure 8.1a) [252] . Furthermore, we investigated publicly available gene expression databases and identified Hs578Ts to highly express SULF1, relative to other breast cancer lines (http://biogps.org/). Hs578Ts are cells derived from human malignant tissue and are exclusively CD44^{hi} [253]. Taken together, we surmise that SULF1 is a marker for cancer stem cells.

Given the high expression of SULF1 in the CSC fraction of various TNBC cell lines, we wanted to investigate if SULF1 plays a functional role in modulating cancer stem cell activity. We knocked SULF1 down using lentiviral delivery of SULF1-targeted

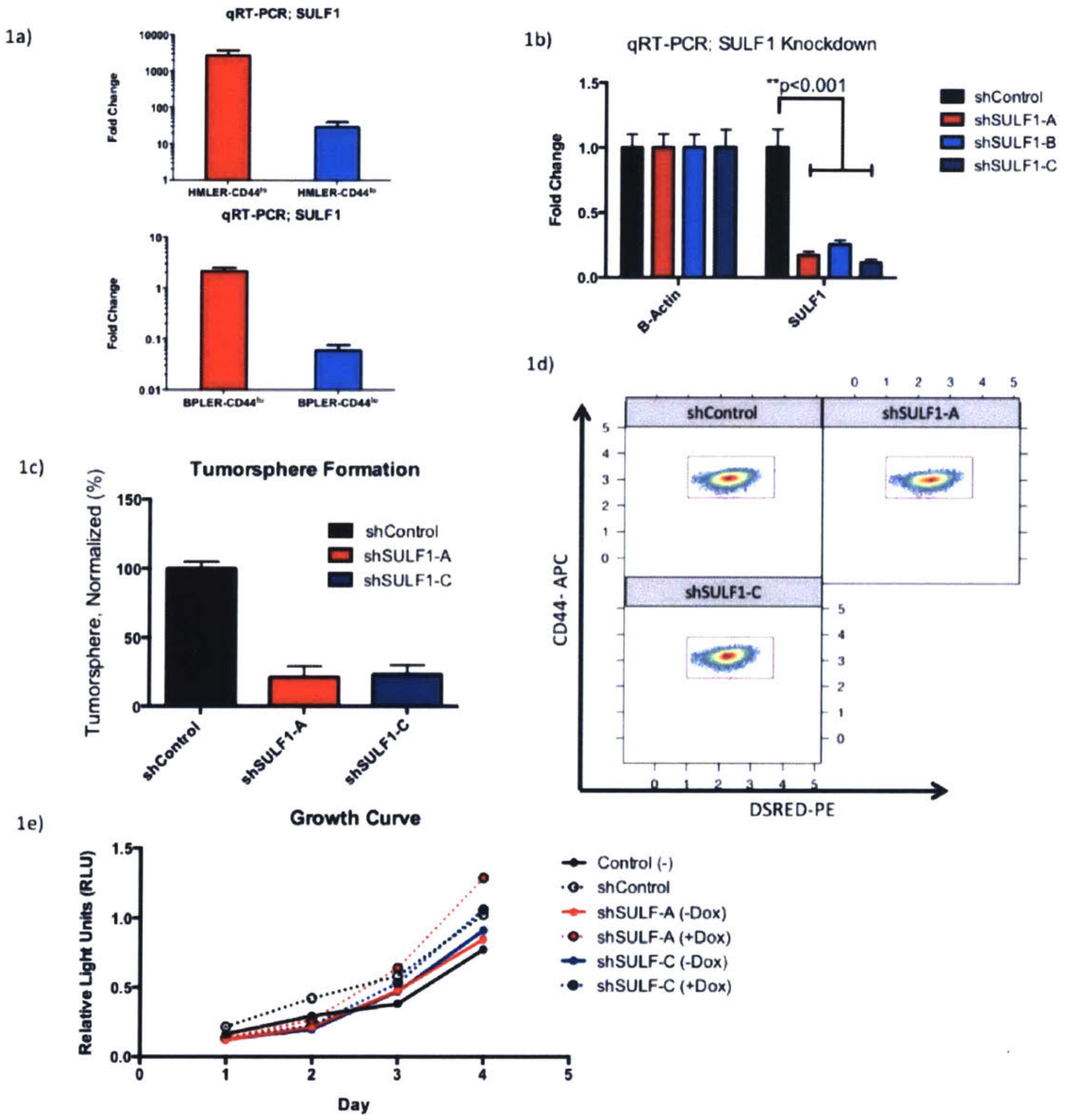


Figure 8.1 **SULF1 is a functional Cancer stem cell marker.** a) qRT-PCR of SULF1 expression in HMLER and BPLER sorted into CD44^{hi} and CD44^{lo} subpopulations. b) qRT-PCR of SULF1 and β-actin in HMLER CD44^{hi} expressing three different stably integrated shRNA Doxycycline inducible constructs targeting SULF1 (denoted shSULF1-A, shSULF1-B, shSULF1-C) compared to the vector only control (shControl). c) Tumorsphere formation assay of HMLER-CD44^{hi} cells expressing SULF1 targeted shRNAs. d) Flow cytometry analysis of the CD44 surface marker in SULF1 knockdown cells compared to the shControl. e) Cell growth curves of control and shSULF1 cells. Data represented here depict the mean ± SEM

shRNA constructs (hereafter referred to as shSULF1-A, -B, & -C or, collectively, shSULF1 cells) in order to assess the functional impact of SULF1 loss on cancer stem cell activity. Using qRT-PCR, we showed that we could reduce SULF1 RNA expression levels by ~75-90%, where shSULF1-C showed the greatest knockdown (Figure 8.1b). Notably, we were unable to assess SULF1 expression at the protein level; numerous commercially available antibody reagents were tested but none showed specificity.

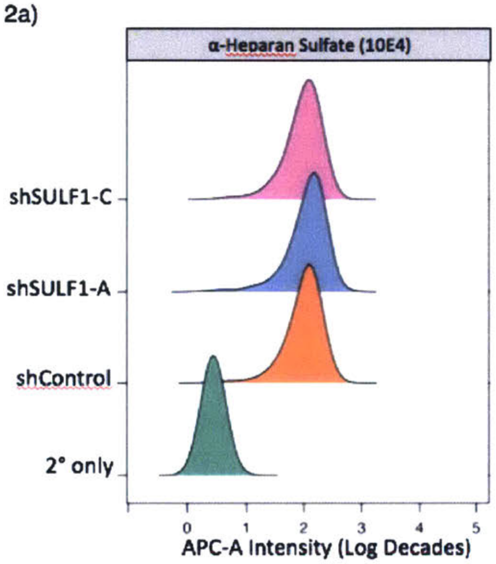
Next, the tumorsphere forming ability was assessed. The tumorsphere assay measures anchorage-independent proliferation of cells seeded at clonal density. The ability to form tumorspheres is a key functional hallmark of cancer stem cells. The HMLER-CD44^{hi} cells form tumorspheres with a frequency of ~10-13 spheres /100 cells. The knockdown of SULF1 however, led to a roughly 70-80% reduction in tumorsphere (Figure 8.1c). It is possible that this phenotype results from some impairment in cell growth properties induced by SULF1 knockdown. After all, SULF1 is known to impinge on mitogenic signaling factors, such as FGF [7]. Thus, we assessed cell growth in 2D cell culture and showed that SULF1 knockdown does not impact cell growth kinetics (Figure 8.1e). Hence, SULF1 is essential for 3D, anchorage independent growth, but not 2D growth.

Another possible explanation for the reduction in tumorsphere formation by the shRNA was that the knockdown of SULF1 pushed the HMLER-CD44^{hi} cells out of their stem-like state into a CD44^{lo} state. As we found, using flow cytometry, this knockdown did not affect their display of the CD44^{hi} marker (Figure 8.1d). Together, these results suggest that the loss of SULF1 in CSCs results in a loss of their tumorsphere-forming powers without affecting their antigen display of the CD44 antigen, an important immunophenotypic marker of CSCs.

SULF1 Modulates the Composition of HSGAGs

SULF1 is an extracellular 6-O-endosulfatase which acts at the cell surface to modify the sulfation of heparan sulfate glycosaminoglycans (HSGAGs), post-synthetically. We aimed to determine if the structure of HSGAGs changed with knockdown of SULF1 expression in the HMLER-CD44^{hi} cells and if so, how the structure was modified. To do so, we employed disaccharide composition analysis, whereby cell-surface HSGAGs are isolated, purified, exhaustively depolymerized using a collection of bacterial derived heparanases (Heparanase I, II, III), and resolved analytically using capillary electrophoresis.

2a)



2b)

	shControl	shSULF1-A	shSULF1-C
Δ-UA(2S)-GlcNS(6S)	2.97	2.14	2.21
Δ-UA(2S)-GlcNS	10.64	9.47	9.35
Δ-UA-GlcNS(6S)	4.93	7.93	5.74
Δ-UA-GlcNS	70.81	64.65	65.21
Δ-UA(2S)-GlcNAc	9.50	8.76	11.69
Δ-UA-GlcNAc(6S)	1.15	7.05	5.81

Figure 8.2 **SULF1 modulates HSGAGs fine structure** a) Quantitation of cell surface heparan sulfate in SULF1 knockdown cells compared to shControl using flow cytometric analysis with an anti-Heparan Sulfate IgM (10E4). b) Disaccharide analysis of Heparan sulfate derived from SULF1 knockdown cells compared to shControl.

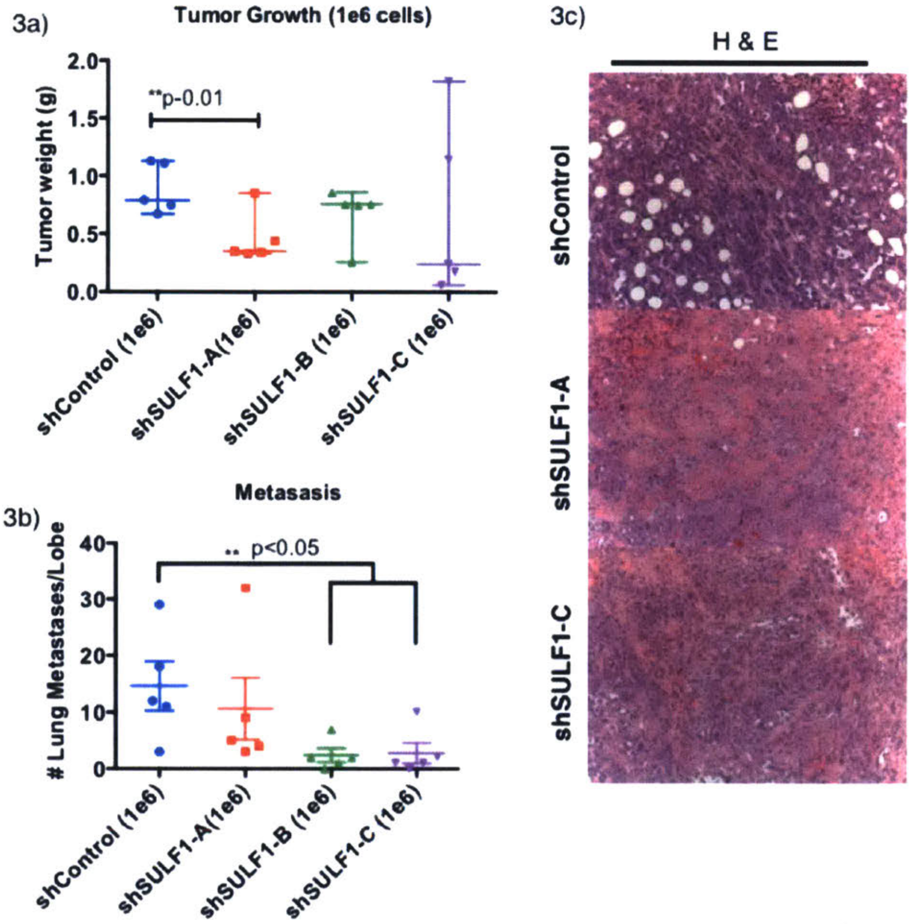
First, we wanted to compare the HSGAG composition on CSC and non-CSCs. To do so, leveraging the framework developed in Chapter 7, we integrated expression of HSGAG biosynthetic enzymes as well as disaccharide composition analysis in HMLER-CD44^{hi} vs HMLER-CD44^{lo} (data not shown). Interestingly, CSCs (relative to HMLER-CD44^{lo}) highly express SULF1, and have a reduced expression of HS6ST1. At the disaccharide level, CSCs had lower abundance of all the 6-O-sulfated disaccharides Δ-UA(2S)-GlcNS(6S), Δ-UA-GlcNS(6S), Δ-UA-GlcNAc(6S), with the tri-sulfated and mono-sulfated 6S disaccharides having a 40% and 50% reduction respectively (data not shown). With this initial observation in hand, we characterized the SULF1 impact on the glycan structure in the knockdown cells.

In the HMLER-CD44^{hi} shSULF1 cells, we observed an increase in two 6-O-sulfate containing disaccharides: Δ-UA-GlcNAc(6S) (~4.9-6 fold increase) and Δ-UA-GlcNS(6S) (~1.2-1.6 fold increase) (Figure 8.2b). Interestingly, we did not observe significant changes in the tri-sulfated 6-O-sulfated disaccharide, which is a reported target of SULF1 [242]. Reduction of SULF1 expression typically results in an increase in Δ-UA(2S)-GlcNS(6S) abundance, however, HSGAG structure results from a complex interplay between biosynthetic enzymes and it is possible other enzymes might compensate. For instance, increases in the tri-sulfated disaccharide driven by loss of SULF1 expression could be negated by compensatory decreases in HS6STs expression, which add 6-O-Sulfation, or by increases in SULF2 expression, which has overlapping activity with SULF1. Notably, SULF2 and HS6ST1 expression was not affected in shSULF1 cells (data not shown). Taken together, our data suggests that in HMLER-CD44^{hi} SULF1's primary role in this context is in the de-sulfation of HSGAG chains containing the 6S mono-sulfated disaccharide, Δ-UA-GlcNAc(6S) (and, Δ -UA-GlcNS(6S), albeit to a lesser degree).

Furthermore, we investigated if the global abundance of HSGAGs had changed after SULF1 knockdown. In order to test this, we used an anti-HS IgM antibody to measure HSGAG abundance by flow cytometry. No significant changes in the abundance of HSGAGs were detected in shSULF1 cells (Figure 8.2a). In summary, we characterized SULF1's effect on HSGAGs in this cell model system. Specifically, we showed shSULF1 cells have an increased abundance of 6S containing dissacharides (Δ-UA-GlcNAc(6S) & Δ -UA-GlcNS(6S)). Additionally, although the knockdown of SULF1 was incomplete and we were unable to measure the SULF1 expression at the protein level, the HSGAG composition analysis provides a surrogate endpoint of SULF1 activity, providing corroborating evidence linking the knockdown level, and the resultant changes in HSGAG structure. Together these analyses show that our shRNA constructs are acting to reduce SULF1 protein levels.

SULF1 regulated tumor initiation, growth and metastasis

Cancer stem cells are functionally defined by their ability to form tumors when implanted at limiting dilutions [254]. In this vein, the molecular signaling networks that regulate tumor initiation and cancer stem cell activity can be assessed similarly. Towards assessing SULF1's function in tumor initiation, thereby validating it as a functional CSC marker, we employed a limiting dilution assay (LDA) measurement to estimate CSC frequency in shSULF1-C vs control cells. Briefly, for the LDA assay, log fold dilutions of cells were implanted into the mammary fatpad of NOD/SCID mice (N=10 mice with $1x10^6$ cells, N=5 with $5x10^5$ cells, N=10 with $1x10^5$ cells, N=5 with $1x10^4$ cells each for shSULF-C & shControl). After 8 weeks, mice were assessed for the presence or absence of tumors and the tumor initiation rate was measured (Figure 8.3d). Using the extreme limiting dilution analysis, an estimation of the number of cancer stem cells was calculated [243]. shSULF1-C cells had an estimated CSC frequency of 1/354,583 (95% CI: 1/643,215-1/195,470) and shControl has an estimated CSC frequency of 1/58,956 (95% CI: 1/121,395-1/28,562) (Figure 8.3d). Thus, loss of SULF1 results in a 5-6 fold decrease in the CSC frequency. Based on high SULF1 expression in the CSCs fraction, and cancer stem cell's dependence on SULF1 to efficiently form tumorspheres & initiate tumors, we conclude that SULF1 is a functional CSC marker, playing an important role in CSC activity.



3d)

	Tumors/Injection		
Cells/injection	shControl	shSULF1-C	Time to Harvest
1x10^{6}	10/10	8/10	8 weeks
5x10^{5}	5/5	1/5	8 weeks
1x10^{5}	7/10	9/10	8 weeks
1x10^{4}	3/5	2/5	8 weeks
Est. CSC Frequency	**1/58,956**	**1/354,583**	P-value =**5.59E-05**
Lower-Upper Est	1/121,395-1/28,562	1/643,215-1/195,470	

Figure 8.3 **Loss of SULF1 reduces tumor growth, metastasis, and tumor initiation** a) Tumorigenicity of SULF1 knockdown cells compared to the shControl following orthotopic transplantation of 1e6 cell into the mammary fat pad of NOD/SCID mice. (n=5/group) Data represented are the median tumor weight ± range. b) Number of GFP+/DSRED+ metastatic lesions in the lung of NOD/SCID mice transplanted with SULF1 knockdown cells compared to shControl. Data represented are the mean number of lung metastases ± SEM. c) H & E stains of parrafinized tumor sections harvested after 8 weeks of growth. d) Tumor initiating frequency and Limiting dilution analysis of SULF1 knockdown cells (shSULF1-C) compared to the shControl.

We also noted that tumors arising from shSULF1 cells had altered characteristics vs control cell. $1x10^6$ cells were mixed with matrigel and injected into the mammary fatpad of NOD/SCID mice (N=5 for each shControl and shSULF1-A, -B, -C), tumor size and number of lung metastases were measured. shSULF1 cells showed a variable tumor growth phenotype and metastatic phenotype (Figure 8.3a and b). As mentioned previously, each construct achieved a variable level of SULF1 knockdown, with shSULF1-C showed RNA levels of SULF1 were reduced ~90%, whereas shSULF1-A and-B showed 85% and 75%, respectively. shSULF1-B and -C showed no statistically significant difference in tumor size versus shControl tumors, but had a reduce propensity to metastasize to the lung (shSULF1-B,–C = ~5mets/lobe vs shControl=~15 metastases/lobe). On the other hand, shSULF1-A showed a significant reduction in tumor growth (shSULF1-A= 0.3g vs shControl =~0.8g), yet it metastasized to the lung with the same frequency as shControl (Figure 8.3a and b). It is possible the variability here is due to the various levels of residual SULF1 protein expression. HSGAGs often exhibit a biphasic dose-response relationship, and they regulate a milieu of biological signals that could differentially influence growth vs metastasis [93,255]. For instance, different levels of residual SULF1, may give rise to different HSGAG structures which ultimately modulate environmental signaling to give rise to two phenotypes: reduced tumor growth while maintaining metastatic potential or normal tumor growth characteristics while losing the ability to metastasize. Unfortunately, due to our inability to quantify the protein level of SULF1 knockdown, we cannot deduce any dose-dependent relationship of SULF1 protein levels to tumor phenotype. Nonetheless, we demonstrate that knockdown of SULF1 expression reduces tumor growth or metastasis. Reviewing the totality of the data, SULF1 plays an important role in the tumor initiation, as well as tumor growth and metastasis.

SULF1 regulates the expression of epithelial and mesenchymal traits

A key challenge emerged throughout the course of this study: the biological phenotypes modulated by SULF1 were only detected in 3D culture or animal model tumors studies. In a way, this is not surprising; HSGAGs, and enzymes that modify

HSGAGs, are functionally critical for interaction between cell-microenvironment (i.e. extracellular matrix, autocrine and paracrine signaling factors) and many biologically relevant interactions regulating tumor initiation are not captured in 2D cell models [256].

In light of these challenges, we wanted to investigate morphology and molecular properties of the shSULF1 tumors. Consistent with previous studies, Hematoxylin-and-eosin (H&E) staining of the shControl tumors showed a morphology consistent with that of an aggressive tumor, with a heterogeneous morphology, containing many fat cells and matrix depositions (Figure 8.3c)[244]. However, tumors arising from shSULF1 cells appear more differentiated, containing epithelial islands, more organized stroma, and have a clear distinction between the stroma and epithelial cells.

To further characterize this phenotype we investigated expression of a tumor differentiation marker, E-Cadherin [257]. Immunofluorescence co-staining using antibodies targeting SV40-LargeT and E-cadherin was performed to assess the tumor

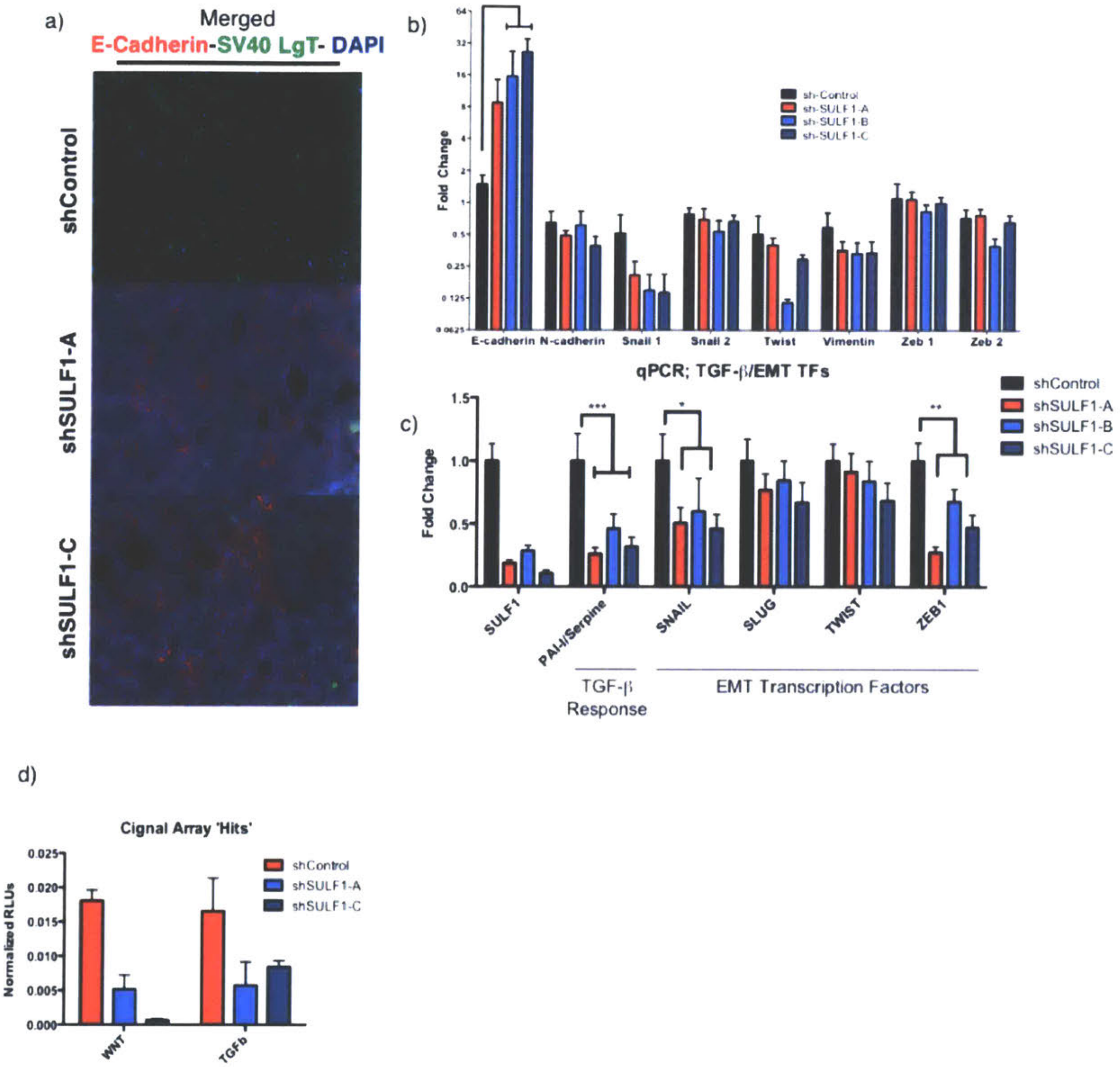


Figure 8.4 **Characterization of the effect of SULF1 knockdown on EMT and signaling.** a) Immunofluorescence of paraffinized tissue sections derived from SULF1 knockdown and shControl tumors (from 3a). Shown here is the merged image of tumor sections co-stained with anti-SV40LrgT (green), anti-E-Cadherin (red), and DAPI (blue). b) qRT-PCR of E-Cadherin and EMT transcription factors on cDNA isolated from SULF1 knockdown and shControl tumors that initiated from 1x10^5 cells injected into the mammary fatpad of NOD/SCID mice. Data are represented as the mean ± SEM c) qRT-PCR of EMT transcription factors and TGFβ response genes on cDNA isolated from SULF1 knockdown and shControl cell lines in 2-D culture. Data are represented as the mean ± SEM. d) Cignal array, luciferase assay based detection of WNT and TGFβ pathway activity.

specific expression of E-cadherin. shControl tumors overall show a low expression of E-cadherin, with only a few pockets of cells that are positive for E-cadherin (Figure 8.4a). In contrast shSULF1 tumors, had a striking increase in the expression of E-cadherin, where the majority of the cells express E-cadherin on the cell surface. Together, these observations confirm that shSULF1 tumors are not only more differentiated histologically, but they also express molecular markers of differentiation.

Molecular components that regulate the epithelial to mesenchymal transition (EMT), are known to play a critical role in tumor differentiation and malignant progression [118]. In fact, E-Cadherin is a marker of the epithelial state, where cells that have undergone an EMT lose the expression of E-Cadherin. Given that shSULF1 tumors, appear more differentiated, we sought to investigate the gene expression of several EMT marker and transcription factors (TFs) in shSULF1 vs control tumors. To determine expression of EMT genes, qRT-PCR analysis was performed on digested tumors (described in methods). The use human specific primers allowed us to only detect the changes in gene expression of the tumors and not the surrounding mouse stroma. A panel of eight genes were used (E-Cadherin, N-Cadherin, Vimentin, SNAIL1, SNAIL2, TWIST, ZEB1, and ZEB2). Cells that undergo an EMT show a loss of expression in E-Cadherin and an increase expression of N-Cadherin and Vimentin[118]. As expected, the differentiation marker E-cadherin was significantly increased compared to the control, which is consistent with the more differentiated morphology of the shSULF1 tumors (Figure 8.4b). Interestingly, the mesenchymal markers vimentin, showed a trend towards decreased expression in shSULF1 cells, but this change was not statistically significant. N-cadherin is not differentially expressed.

Next, the expression of five different EMT transcription factors, responsible for EMT induction, were analyzed (SNAIL1, SNAIL2, TWIST, ZEB1, and ZEB2). Only SNAIL1 expression was decreased in shSULF1 tumors relative to shControl. The other transcription factors were not differentially expressed (Figure 8.4b). This finding is in line with the current knowledge on regulation of various epithelial and mesenchymal genes by EMT-TFs. SNAIL1 represses E-Cadherin expression, whereas other TFs like ZEB1 are predominantly responsible for activating expression of mesenchymal markers

(Vimentin/N-cadherin)[121]. Based on these results we surmise, that reduction in SNAIL1 expression enables a gain of epithelial features (i.e. expression of E-Cadherin), but not a loss of mesenchymal features.

Next, we investigated the expression of these EMT-TFs in 2D culture to assess if these trends held true outside of the tumor tissue context. Interestingly, both SNAIL1 and Zeb1 expression was decreased in shSULF1 vs shControl cells, while SNAIL2 (also known as SLUG) and TWIST were not differentially expressed (Figure 8.4c).

Together these findings confirmed our initial observations that SULF1 knockdown tumors are more differentiated compared to the shControl tumors. Additionally, the evidence here suggests that SULF1 knockdown results in a gain of epithelial features coupled with the decreased expression in EMT-TFs (SNAIL1 in tumors, SNAIL1 and ZEB1 in cultured cells). These results can be interpreted as a loss of the EMT'd state, or as a mesenchymal to epithelial transition (MET). Nonetheless, our work here shows that SULF1 may play a role in the maintenance and regulation of EMT. To our knowledge, no such role has been reported for SULF1 in breast cancer.

Next, our goal was to uncover a possible mechanism by which SULF1 might regulate the cancer stem cell activity, through its regulation of the EMT. Previous work performed in Bob Weinberg's lab uncovered that paracrine and autocrine signaling factors, WNT and TGFβ, play a critical role in the induction and maintenance of EMT in both stem cells and cancer stem cells[69]. Importantly, both of these factors are regulated by heparin sulfate and specifically by SULF1. In the case of both WNT and TGFβ signaling, their activity is negatively regulated by 6-O-Sulfation of HSGAGs, therefore knockdown of SULF1 expression should decrease their signaling activity . Thus, we investigated if WNT or TGFβ signaling was altered after the knockdown of SULF1. To do so, the expression of a TGFβ response gene, PAI-I/Serpine was measured using qRT-PCR and found to be significantly reduced in shSULF1 cells (Figure 8.4c). Furthermore, a luciferase-based signaling assay was performed to investigate activity of the WNT and TGFβ pathways (Cignal reporter array, CCA-001L; SA biosciences). Importantly, both WNT and TGFβ signaling activity was significantly reduced in shSULF1 cells compared to shControl(Figure 8.4d). Both TGFβ and WNT are known to promote EMT through upregulation of SNAIL1. Furthermore, both signals are important for the maintenance of EMT. Additionally, inhibition of both signals has been reported to play a role in the mesenchymal to epithelial transition. Taken together, a likely mechanism explaining the more differentiated nature of shSULF1 tumors is that a knockdown of SULF1 decreases WNT and TGFb signaling, resulting in a decreased SNAIL1 expression, thereby relieving transcriptional repression on E-Cadherin.

SULF1 is upregulated in human breast cancer and is negatively correlated with relapse free survival

Given that knockdown of SULF1 results in a decrease in CSC activity, likely driven by an MET, SULF1 might be an interesting therapeutic target. Towards understanding the translational relevance of our finding to human tumor biology, we investigated the expression of SULF1 in patient tumors. Using publically available gene expression databases (www.Oncomine.org; TCGA data), we assessed expression of SULF1 in invasive breast carcinoma vs matched normal healthy tissue [258]. These results show that SULF1 is expressed ~3.4 fold higher in invasive breast carcinoma vs normal tissue (Figure 8.5a).

Our work suggests that SULF1 is a viable therapeutic target, and inhibition of SULF1 could reduce CSC activity in triple negative breast cancer. Given that SULF1 is highly expressed in invasive breast cancer, we wanted to understand if SULF1 expression correlated with patient outcomes, in particular in cases of triple negative breast cancer. Using Kmplot, we investigated correlation between SULF1 expression and relapse free survival in ER-/PR- Basal BrCa (which is the databases closest correlate to TNBC)[259]. Indeed, high SULF1 expression correlated with poorer relapse free survival (HR= 2.05) (Figure 8.5b). These results, while only correlative, provide some validation that SULF1 is a clinically relevant player in breast tumor progression, and TNBC in particular.

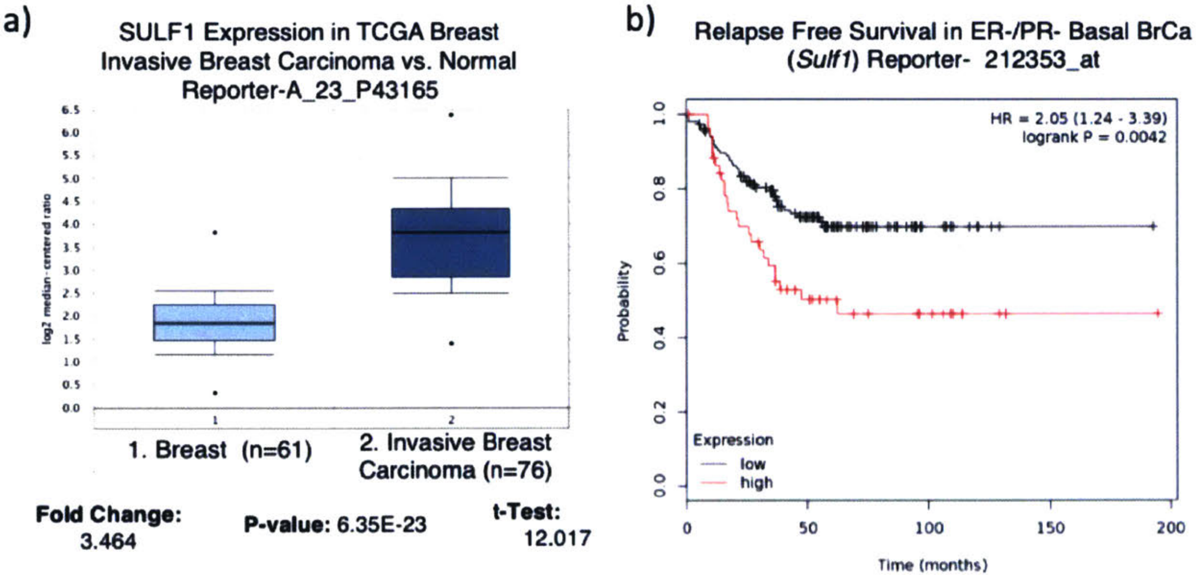


Figure 8.5 **SULF1 is upregulated in Invasive Breast Carcinoma and high expression is associated with a poor clinical outcome in ER/PR- Basal Breast Cancer.** a) SULF1 gene expression in normal breast compared to patients with Invasive Breast Carcinoma. The plot shown here is derived from the TCGA breast gene expression dataset on the Oncomine database (www.oncomine.org). Data are presented as box and whisker plots ± range. b) Kaplan Meier plot depicting Relapse Free Survival of ER-/PR- Basal Breast Cancer patients comparing a high SULF1 expressing cohort with a low SULF1 expressing cohort. The data represented here was derived from Kaplan-Meier plotter breast cancer resource (www.kmplot.com).

Conclusion

Here, we conduct the first investigation into how an HSGAG-modifying enzyme, SULF1, affects breast CSC behavior through modulation of cell-surface HSGAG sulfation. We show that SULF1 is highly upregulated in the CSC fraction of several TNBC cell models. Furthermore, we report that knockdown of SULF1 in a TNBC CSC model reduces tumor initiation, growth and metastasis and gives rise to histologically more well-differentiated tumors that more strongly express the epithelial marker, E-Cadherin. Additionally, in tumors we find that knockdown of SULF1 decreases the expression of the EMT-TF SNAIL1, suggesting that knockdown of SULF1 may promote a mesenchymal to epithelial transition (MET). We further suggest that the mechanism by which knockdown of SULF1 drives MET is through the HSGAG specific regulation of WNT and TGFβ.

We envision a mechanism whereby loss of SULF1 expression in cancer stem cells results in an increase in 6-O Sulfation on HSGAGs. The changes to the structure of HSGAGs (i.e. the increase in 6-O-sulfated disaccharides) creates a high affinity binding sites for WNT ligands acting to sequester WNT from its receptor Frizzled, thereby reducing signaling activity [88]. A similar mechanism for TGFβ1 has been reported, where it preferentially binds to 6-O Sulfated HSGAGs on TGFβR3, and in the presence of SULF1 is liberated and can interact with its co-receptor[260]. Importantly, both WNT and TGFβ1 signaling are known to act to maintain the expression of SNAIL1, leading to suppression of epithelial characteristics in CSCs[69,121]. In the case of SULF1 knockdown, SNAIL1 expression is reduced, leading to CSCs acquisition of epithelial characteristic.

The results presented here suggest that SULF1 is critical for the extracellular signals that are responsible for maintenance of EMT signaling and CSC behavior. To our knowledge, this is the first report demonstrating the role of SULF1 in breast cancer stem cells, as well as maintenance of EMT in breast cancer stem cells.

As mentioned above, the CSCs that exist in tumors are likely candidates for the initiation of tumors and metastatic outgrowth. For these reasons it is imperative that therapies specifically targeting the biology of CSCs are derived. Ideally, such therapies should be used in conjunction with conventional therapies to eradicate CSCs within mammary carcinomas. In this way, one should be able to eradicate the CSC populations that drive cancer progression and thereby inhibit cancer recurrence and metastasis, and ultimately, improve patient survival. Thus, the identification of SULF1 and its essential role in CSC physiology and regulation of the EMT suggests future attempts at targeting this enzyme may enhance the susceptibility of CSCs to therapeutic intervention.

Acknowledgements

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Bob Weinberg, Christine Chaffer, Rahul Raman, Oliver Dodd, Susanne Janssen, Ference Reinhardt, Brian Birie and Ram Sasisekharan for their contributions to this work. This work was funded in part by the Koch Institute Frontier Research Award, National Institutes of Health (T32 ES007020-38) and MIT UROP Funding Office.

Chapter 9 Enabling the development of drugs derived from farm dust extracts

Summary and Significance

Over the last three decades, an interesting phenomenon has emerged demonstrating that exposure to certain environments with high microbial exposures, such as the traditional farm environment, can protect against asthma and atopy. The strength and robust nature of this protective effect have prompted numerous inquiries into the mechanisms that give rise to this phenomenon. Understanding the causative environmental factors and immunobiology of this effect could yield novel insight into therapeutic strategies to treat asthma or allergies. New evidence suggests that the protective effect can be conferred through exposure to dust isolated from cowsheds. In fact, extracts from cowshed dust are protective in animal models of asthma. Dust extracts are complex, heterogeneous mixtures and comprised of various bioactive molecules, including complex polysaccharides. While the exact active ingredient is unknown, one viable therapeutic strategy is to use the crude dust extract itself. One challenge with this approach is that developing therapeutics from complex mixtures, where the active ingredient (API) is poorly defined, poses a significant regulatory challenge, specifically concerning safety and consistent potency. Approved therapeutics such as Heparin or Glatiramer acetate faced similar challenges and in these instances, integrated analytical tools have played a crucial role in defining the "product" and ensuring quality, consistency, and potency.

In chapters 7 and 8 of this thesis, I implemented integrated analytical strategies to decode glycan structure-function relationships. Drawing on this expertise, in this chapter, I apply a similar framework and toolset to conceptualize an integrated analytical strategy to decode composition-activity relationships of extracts. Furthermore, I outline an analytical workflow to assess product consistency and potency for therapeutic farm dust extracts. Together, the work here provides a regulatory path enabling the development of farm dust extract-derived therapeutics.

Introduction

The prevalence of allergic-type disease, such as asthma and allergies, has been on the rise in western countries since the 1960s[261]. Currently, allergic-type diseases represent a major health crisis. Asthma alone affects over 278 million people worldwide [262]. There is, thus, a great need to develop preventative strategies and new therapies. One powerful method of identifying novel therapeutic strategies is to identify populations with naturally lower prevalence of allergic-type diseases and study the environmental exposures and mechanisms that underlie their protection.

Recently, an interesting phenomenon has emerged demonstrating that exposure to certain environments with high microbial exposures can protect against allergic-type diseases. This finding has been explained by the "hygiene hypothesis", which posits that allergic-type disease susceptibility is increased in populations that have a reduced exposure to microorganisms, both in quantity and diversity[263]. This hypothesis has spawned many well-powered epidemiological studies investigating relationships between the prevalence of allergic-type diseases and exposure to "dirty" environments. The traditional farm environment represents one example of an environment of high microbial exposures, which is conveyed by close contact to farm animals and their fodder and bedding (straw and hay).

There is robust evidence demonstrating that farm exposure has a potent protective effect against asthma and atopy[261]. Over 30 studies have demonstrated this protective 'farm effect' in various European countries, New Zealand, Australia and the US[1]. These findings have been substantiated in animal models of allergic asthma where aqueous extracts of farm dust showed a protective effect[264–266]. Another example of this was recently published in the New England Journal of Medicine (NEJM) where asthma risk was compared between Amish and Hutterite farm children[265]. These two populations have similar genetic ancestry and lifestyle practices, with the largest differences being in their respective farming practices. The Amish employ traditional farming practices using horses instead of machines for work, while Hutterites use industrialized practices. In these two populations, asthma prevalence and allergic sensitization was 4-6 times lower in the Amish cohort. Airborne and settled dust collected indoors in the Amish households showed a higher concentration of endotoxin and common allergens compared to Hutterite indoor dust. When administered prophylactically in a mouse model of allergic asthma, Amish dust extract protected against asthma whereas Hutterite dust extract exacerbated disease[265]. Together these studies suggest that certain components present in farm dust are likely responsible for the protective effect observed in children raised on farms.

Given the aforementioned results, if one were able to analytically characterize components in this mixture or, perhaps, isolate the active molecule(s) in farm dust which are responsible for the protective effect, then prophylactics for asthma and allergies can be developed. This manuscript will summarize the epidemiological evidence linking farm dust exposure to protection from allergy and asthma as well as the evidence suggesting that components in farm dust are responsible for the protective effect. Furthermore, I will propose a framework for addressing regulatory challenges associated with developing novel farm dust extract therapeutic products, specifically focusing on methods for characterizing farm dust extract analytically and methods of testing biological activity.

Epidemiological evidence supporting the "farm effect"

The "hygiene hypothesis" was coined in 1989 by David Strachan in his search for identifying factors that caused the increase in prevalence of hay fever, asthma, and childhood eczema that followed the industrial revolution[267]. In his study, he included 17,414 British children born during one week in March, 1958, and followed them until age 23. He noted that a larger household size was inversely correlated with prevalence of hay fever. Strachan hypothesized that allergic disease was prevented by repeated infections during childhood, which occurred with increased contact with unhygienic older siblings. He further suggested that the industrial revolution brought higher standards of hygiene, particularly in wealthy families. These cleanliness habits reduced exposure to pathogens and resulted in increased prevalence of hay fever[267]. Since Strachan's coining of the hygiene hypothesis, numerous high-quality epidemiological studied have been conducted with the goal of exploring this effect.

Farm-dwelling people are unique populations that have been instrumental in elucidating this effect. It has long been observed that children of farmers have a reduced prevalence of allergies and asthma[268–270]. This observation fits with Strachan's hygiene hypothesis, as a farm lifestyle typically includes large sibship size, exposure to plant material, interaction with livestock and their products (manure, milk), and use of indoor coal and wood burning stoves which all result in an increase in the diversity of numerous microbial exposures. Several seminal epidemiological studies sought to characterize the farm effect in depth addressing questions surrounding strength of the effect and what factors specific to the farm lifestyle contribute to this effect[269,271–273]. Other reviews and meta-analyses have summarized the full scope of epidemiological data[261,274]. For the sake of brevity, I will discuss what I consider to be the key epidemiological features of the farm effect focusing on asthma and atopy. Specifically, I will discuss exposure-response relationships, timing of exposure, specific farm factors which are likely responsible for the effect, and the immunobiology of the effect.

Certain farm exposures have consistently been identified across several studies as contributing to the reduced risk of asthma and allergies: contact with livestock, mostly cattle, contact with animal feed such as hay and silage and the consumption of raw cow's milk[268,271]. These exposures had an independent protective effect after mutual corrections in multivariate models. Other differences in lifestyle such as duration of breast feeding, family size, pet ownership, other dietary habits, parental education or a family history of asthma and allergies did not account for the protective 'farm effect'.

The protective effect of a farming lifestyle appears to be dependent on the timing of the initial exposure, and the frequency of exposure. The cross-sectional ALlergy and EndotoXin (ALEX) study demonstrated that children exposed to animal sheds within the first year of life had a stronger protective effect than those who were exposed after the first year of life[269]. Still, those exposed after the first year of life still showed some protective effect relative to those who were not exposed at all. In the prospective farm birth cohort, the PASTURE study, exposure to animal sheds in the first year of life was associated with a significantly reduced risk of wheezing (aOR=0.44, 95%CI:0.33-0.60).

Thus, exposure to the farm environment very early in life (<3 years old) may shape the immune system of these children, resulting in sustained immunological changes that protect against allergic type diseases[269,275–278].

These findings also suggest that exposures encountered in animal sheds, in particular cattle stables, play a major role. It has been known for decades that a large variety of bacteria, fungi and their compounds are abundant in animal sheds[279]. Children play in and around these animal sheds and take their microbial exposures into the indoor environment where micro-organisms and their compounds settle in floor and mattress dust. Recent work has, in fact, shown a high concordance between microbial flora of cowshed dusts and farm children's mattress dust[280]. Mattress dust can thus be regarded as a reservoir reflecting a subject's long term exposure encountered in indoor and outdoor environments.

The level of microbial exposure from mattress dust has been assessed in a number of studies by measuring markers of bacterial and fungal exposures. Muramic acid which is a component of peptidoglycan, is a cell wall component of all bacteria, but more abundant in Gram-positive bacteria. Endotoxin (LPS) is a cell wall component of Gram-negative bacteria only. Extracellular polysaccharides (EPS) are derived from Penicillium spp. and Aspergillus spp. and are thus a marker of fungal exposures. Levels of markers of bacterial and fungal exposures have been shown to be inversely related to asthma and atopy in farm and also in non-farm environments, in children as well as adults[272,281,282]. These associations were strong; the prevalence of atopy reduced more than 10 to 25 percent with increasing microbial exposure. These findings therefore support the notion that microbial exposures may account at least in part for the protective 'farm effect'.

Immunobiology of the "Farm effect"

In order to develop therapeutic or prophylactic strategies, understanding the immunological mechanisms that drive and maintain this protective effect is critical. Studies which have tried to elucidate the mechanism of immune alteration, have shown that the farming lifestyle influences components of both the innate and the adaptive immune system. The mechanisms have been reviewed in detail elsewhere, so I will only briefly summarize a few of the most well-characterized effects[261].

Pathogen-associated molecular patterns (PAMPs), which are highly conserved structural components of microbes, are recognized by pattern recognition receptors (PRRs), which are similarly conserved receptors present in host innate immune systems. Examples for PAMPs include: Extracellular polysaccharides (EPS), endotoxin (lipopolysaccharide (LPS)) and peptidoglycan. In humans, examples of PRRs include Toll-like receptors (TLRs) and CD14. At present, ten functional TLRs have been described in humans. The cellular signaling cascade following engagement of TLRs is responsible for initiating innate host defense mechanisms[283], and providing signals required for initiating and modulating the adaptive immune response [284]. If environmental microbial exposure affects the development of asthma and allergies, then pattern recognition receptors should be involved in the pathogenetic process. Gene expression of TLRs and CD14 was studied in farm populations. Peripheral blood leukocytes from children of the ALEX population living on a farm were found to display increased expression in a subset of PRR genes encoding CD14, TLR2, and TLR4 as compared to non-farm children[285]. The impact of farming on the expression of innate immunity genes was then further examined in the PARSIFAL study[278]. A number of individual farm characteristics were related to the upregulation of distinct TLR genes[272]. To provide proper homeostasis of the innate immune response, a complex regulatory network has evolved, including downstream adaptor proteins such as MYD88 and TRIFF. It is noteworthy that mice lacking these adaptor proteins in knockout mutants were no longer protected by Amish dust extracts[265]. These findings add biological plausibility to the epidemiological findings and also suggest underlying mechanisms, i.e. an activation of the innate immune response via activation of pattern recognition receptors.

A recent study published in the New England Journal of Medicine supports this notion by showing that the asthma protective effect specific to the Amish farming lifestyle (versus Hutterites), was correlated with changes in both proportion and phenotype of innate immune cells. The peripheral blood leukocyte (PBL) isolated from

Amish children showed an increased proportion of neutrophils and decreased eosinophils[265], relative to Hutterite children. Furthermore, the Amish children's neutrophils showed an immature phenotype, consistent with recent or frequent exposure to microbes, and their monocytes showed a suppressive phenotype, with decreased HLA-DR and increasing ILT3 expression. The gene expression profile of PBLs also showed increased expression of genes related to innate immunity[265]. Taken together, the measurements presented here provide a roadmap for a protective immunophenotype in innate immune cells. This is enabling in the context of drug development, as I have putative biomarkers that one could measure during clinical trials to monitor whether or not a protective effect has been achieved by a farm dust extract therapy.

Farm-dust-derived bioactive molecules

The epidemiological evidence suggests that there are molecular agents that humans are exposed to which are responsible for the protective effect. This has been substantiated using several animal and cell models of asthma and allergies, where potential causative agents, including aqueous stable dust extracts, microbes or microbial constituents, have been tested for efficacy[264–266,286]. Such systems allow us to explore how various constituents that people are exposed to on the farm may modulate the immune system.

The first study that demonstrated that it is possible to experimentally isolate immunomodulatory substances from farms was published in 2006[266]. Aqueous extracts were created from dust collected from cattle stables from the ALEX study. The sediment dust was collected from surfaces between 0.05-0.15m above the ground, insinuating that settled dust on surfaces above the ground were in the air at one point, leaving the possibility of airborne exposure. The dust extract was administered in aerosol form, prophylactically, to an OVA-model of allergic asthma. The farm dust exposure resulted in dramatically reduced airway hyper-responsiveness (AHR) to methacholine provocation, and also suppressed eosinophilia, as measured by cell composition in bronchoalveolar lavage. This study represents two advancements: first, it empirically established the presence of immunomodulatory biomolecules in farm dust and second, it established a well-characterized model system to identify biological relevant compounds present in farm dust extracts. Importantly, dust extracts sourced from various farm environments exist and show differential activity. For instance, dust from raised surfaces of central European cattle sheds, and dust collected electrostatically from Amish households are protective against asthma, while dust isolated from Hutterite households are not[265,266].

Next, a number of bacterial species occurring in cowsheds of farms were identified by direct culture. Two isolates were selected because of their relative abundance in cowshed microflora and the farm children's IgG and IgA antibody responses to these species, namely *Acinetobacter lwoffii* F78 (Gram-negative) and *Lactococcus lactis* G121 (Gram-positive) were examined[286]. The asthma and allergy preventive potential of both strains was investigated using the same animal model. BALB/c mice were treated, intranasally, with 10^8 cfu of lyophilized bacteria beginning 10 days before sensitization and continuing over the sensitization and challenge process. Exposure to either bacterial strain markedly suppressed allergic airway inflammation as assessed by eosinophils in the BAL fluid, lung histology and lung function. Additional *in vitro* studies showed that both strains induced a Th1-polarizing program in dendritic cells [286]. Therefore, exposure to these microbes originating from farming environments may induce Th1 immune responses which may counterbalance the asthma and allergy inducing Th2 responses.

Currently, no extensive analytical efforts have been undertaken to deeply identify compositions of dust extracts; therefore, it is possible that there are additional components in the dust. These factors could include a mixture of components from bacteria or fungi (both viable or non-viable components), and components derived from animals or plants. Additionally, because of the complexity of the sample mixture, it is possible that multiple components in the farm dust extracts, and not any one component or class of components can give rise to the long lasting protective effect. Therefore, in order to develop therapies that are derived from farm dust, I need to identify a suite of orthogonal measurements that will allow us to analytically define the composition of bioactive farm dust.

From dust to drug: A regulatory framework to characterize farm dust extract

The epidemiological data and animal model studies summarized above suggest that dust extracts from certain traditional farming environments could be used to prevent asthma. One simple incarnation of a potential therapeutic product, could be an extract of dust collected from an environment known to be protective, such as Amish houses or raised surfaces in cowshed in central Europe (hereafter referred to as protective dust extracts[1]). From a regulatory perspective, this poses a significant challenge: how does one develop a therapeutic from a complex mixture where the active ingredient is hard to unequivocally define? Inspiration can be drawn from guidance on botanical drugs or certain cases studies of biosimilar drugs where analytical tools have played a crucial role in defining the "product" and ensuring quality, consistency, and potency[287–290].

[1] The dust extract(s) I will consider here are prepared using methodology described previously[266,291].

The development of a protective dust extract depends on the development of an appropriate analytical framework. The major challenge with a dust extract therapeutic product is that there is no known active ingredient, and there is likely a high degree of variability that arises during sourcing and production. It is important to ensure that the product that patients will be receive is free from harmful substances, and is consistent in its efficacy and potency. The development of an analytical framework can address both of these challenges. An appropriate analytical workflow should both qualitatively and quantitatively capture the composition of the drug substance. It should incorporate the measurement of molecules known to be present in the mixture, crude compositional analysis (i.e. DNA, amino acid, or lipid content), as well as fingerprinting strategies. Furthermore, it should incorporate appropriate biological activity to assess potency. Herein I will outline a putative analytical workflow, which incorporates the above criteria. I do not make claims as to which specific technologies are optimal, but rather state the high-level objectives of such measurements and representative examples of technologies.

First, it is important to acknowledge several important challenges associated with performing analytics on dust extracts. Protective dust extracts are comprised of DNA, lipids, proteins, and carbohydrates. No single tool or technology is capable of simultaneously measuring all of the diverse classes of analytes. As such, an appropriate analytical workflow requires multiple integrated analytical methodologies to capture all of these diverse analyses. A second major challenge is that the biomolecules in dust come from a variety of organisms including plants, animals, microbes, fungi and insects. This poses significant challenges in the implementation of methods and the interpretation of data from 'omics-type measurements. Proteins derived from different organisms will vary in their post-translational modifications, which makes interpreting peptide sequencing and MS fragmentation data challenging. Similarly, analysis of carbohydrates requires prior knowledge of the organisms of origin because carbohydrates possess such a high degree of organism specific variation in structure (i.e. linkage or arrangement) and composition (i.e. types of monosaccharide building blocks). Given these challenges, it is not feasible to use 'omics-type approaches with the goal of unambiguously identifying specific analyte structure or sequences. As such, an appropriate analytical framework should focus on employing robust compositional measurements and fingerprinting.

Crude compositional, or building block, measurements have been used extensively in complex mixtures analysis. Product measurements should include percent-by-weight of DNA, proteins, lipids, simple and complex carbohydrates, as well as relative abundances of individual amino acids, fatty acids, and carbohydrate building blocks. These types of measurements are robust in face of the challenges introduced by the organismal and species diversity. Similar measurements have proven useful in cases of demonstrating active ingredient sameness for generic glatiramer acetate[2]. For instance, the peptides comprising glatiramer acetate are depolymerized in order to measure the molar fraction of amino acids (Ala, Glu, Lys, Tyr); this allowed manufacturers to show sameness between the generic (Glatopa, Sandoz) and branded drug (Copaxone, Teva) [288]. Importantly, some crude composition analyses (carbohydrate and lipid analyses) have been performed on protective dust extracts previously[291], and they were effective in elucidating the product-process relationships in farm-dust extracts.

Fingerprinting with high sensitivity analytical tools provides important complementary data sets to the simple compositional analysis described above. Tools such as SDS-PAGE, liquid chromatography (i.e. SEC or IEX HPLC), spectroscopy (i.e. ATR-FTIR) and mass spectrometry (i.e.MALDI-MS) could provide signatures or characteristic readouts (i.e. retention time of specific peaks on HPLC column) for a protective dust extract. These measures could be useful in capturing batch-to-batch variation. Furthermore, integrating such analytical tools with enzyme treatments or chemical extraction provide meaningful, complementary information. These methods have proven useful in determining active ingredient sameness of generic

[2] Glatiramer acetate (Copaxone (Teva Pharmaceuticals USA Inc., NorthWales, PA, USA) is a random polymer of four amino acids found in myelin basic protein (glutamic acid, lysine, alanine and tyrosine) and is approved for the treatment of relapsing forms of multiple sclerosis

enoxaparin[289][3] to the innovator drug. In this example, multiple high-resolution analytical methodologies were used including: structural signatures of heparins or heparin preparations obtained using NMR[292] and de-polymerization with enzymes followed by HPLC and mass spectrometry analysis of intermediate products[293].

In addition to composition and fingerprinting measurements, it is will be important to quantify the abundance of known constituents. As mentioned earlier, there are numerous compounds that have been identified in farm dust. Two commonly identified substances are endotoxin, and muramic acid[281,282]. These can be measured using kinetic limulus assays and gas-liquid chromatography/mass spectrometry analysis, respectively, and are considered to be markers of microbial exposures. Importantly, given that these two substances are consistently present in most dust extract samples they provide a targeted way to capture batch-to-batch variation in dust. Other constituents that are present include allergens, like Der p 1, arabinogalactans, and various microbes or spores.

Bioactivity assays are also a critical part of an integrated analytical framework. In the cases of generic enoxaparin and generic glatiramer acetate, both leverage multiple, and often redundant, orthogonal bioactivity assays to assess the potency and active ingredient sameness[288,289]. In the case of generic enoxaparin, the “sameness” evaluation includes assessment anti-coagulant properties, like biochemical assays measuring factor Xa and factor IIa inhibitory activity[289]. Additionally, equivalence must further be demonstrated by an in vivo pharmacodynamics profile measuring anti-Xa and anti-IIa activities in healthy volunteers[289]. In the case of biosimilar glatiramer acetate (GA) showing active ingredient sameness depended on measuring GAs effect on i) functional activity of APCs, B-cells and T-cells, ii) changes to genome-wide gene expression and iii) *in vivo* activity on clinical scores of multiple animal models of experimental autoimmune encephalitis[288]. Similarly, measuring potency of protective

[3] Enoxaparin belongs to a class of drugs known as low-molecular-weight heparins. It is an injectable product and is used to prevent deep vein thrombosis, which is a blood clot in a deep vein that may lead to pulmonary embolism.

dust extracts using biological activity assays will be critical in monitoring batch-to-batch variation. An appropriate scheme requires integrating measurement of a dust extracts effect on genome-wide expression changes, cell signaling and activity assays across relevant cell model systems of lung tissue and immune cells. Additionally, it will be important to include assays measuring *in vivo* activities of protective dust extracts. OVA and HDM models of allergic asthma have been used previously the efficacy of farm dust extracts[264–266]. Importantly, one would need to include key read outs in these model system including: Airway responses and bronchoalveolar-lavage cellularity following methacholine or acetylcholine challenge and OVA or HDM specific IgE and IgG2a titers.

Efforts should also be undertaken to test for impurities or other substances that might be harmful to humans. Given that most of the protective dust extracts are isolated from traditional farming environments, the guidance on botanicals can be applied here. Extracts should be tested for presence of residual pesticides or antibiotics used on the farm and in barns, elemental impurities (i.e. heavy metals), known pathogens, and adventitious toxins[287]. It will also be important to quantify various allergens, because it is possible that the presence of certain allergens might preclude a patient with an existing allergy from receiving a protective dust extract product.

Other final key element related to the regulatory aspects of commercializing a protective dust extracts product is clinical trials design. Due to the similarities in dust extract products and botanicals, the FDA's botanical guidance provides a reasonable framework for design and execution of clinical trials; I will not discuss this here. It is worth considering, from a human safety perspective, that there is documented evidence of adults, children and pregnant women having daily exposure to dust in protective environments. Given this, it is possible to parameterize both average and maximum daily exposure of dust intake, which can be used to estimate safe dosing limits of the therapeutic. To estimate the correct dose for human studies, I recommend a two pronged approach: i) estimating current daily exposure levels in persons inhabiting the protective farm environments and ii) animal model dose-finding studies.

Conclusion

In this review, I summarize the epidemiological evidence demonstrating the traditional farm environments' protective effects on asthma and allergy. I further outline which environmental factors most likely responsible for the effect, and present some mechanistic insights into the immunobiology of the protectives effect. Together these support the idea that there is potential for a robust therapeutic intervention derived from the farm environment. An extract made using dust obtained from these farm environments could be a powerful therapeutic, but because the bioactive ingredient has not yet been identified strategies are needed to overcome the regulatory challenges associated with development. The development of a farm-dust extract is possible with the implementation of an analytical framework. The major goals of our analytical frameworks are to ensure that farm dust extracts are safe and consistent in their composition and potency. Using an integrated analytical approach comprised of compositional analysis, high-resolution analytical fingerprinting assays and a suite of biological activity assays it is possible to achieve these goals, thereby enabling the development of farm dust extracts as a therapeutic to prevent or treat asthma.

It is possible that instead of using the simple extract describe above[a], one might be able to identify production/ strategies that enrich its activity or identify active ingredient(s) present in dust extracts which are responsible for their protective effects. Recently, this type of approach was used to harness the autoimmune disease inhibitory activity of hookworm parasites[294]. It has been known for many years that populations with a high prevalence of hookworm infections have strikingly lower prevalence of allergies or autoimmune disease. In fact, clinical trials have been performed investigating the effects that hookworms had upon infection of patients with autoimmune diseases[294,295]. Unfortunately, there are numerous challenges associated with this approach: there is heterogeneity in biological activity of hookworms, hookworms are hard to cultivate at scale, and infecting patients with live parasites poses potential health risks. Because of these challenges, scientists recently undertook efforts to identify immunomodulatory proteins present in the oral secretions of hookworms. They identified a protein, called AIP-2, present in oral secretions which could provide potent and long lasting protective immunomodulatory activity in animal models of asthma when administered prophylactically and therapeutically[296]. One could apply a similar approach to farm dust extracts. Importantly, efforts to build out composition-activity relationships require two key parts: First, it requires the same analytical pipeline I presented above and, second, it requires systematic strategy to perturb the composition of farm dust (i.e. using chemical enrichment strategies, enzymatic treatments, or chromatography fractionation by size or charge). While identifying a single protein or molecules sounds like an ideal strategy, it's important to note that it may not be possible to distill the activity down into one or several single biomolecules as the observed clinical outcomes might be dependent on the mixtures' complexity. At this point, the latter remains unknown.

The work outlined here can also be leveraged to address challenges emerging in new medicines that are derived from botanicals, microbes, and even traditional Chinese medicines (TCMs). Inevitably, many such medicines will contain active ingredients which may be therapeutically useful, but hard to unequivocally define and, therefore, may challenge the development of products with robust safety and efficacy profiles. I hope the work presented above provides a regulatory framework that helps accelerate the development of promising therapeutic candidates that might have been previously limited because of their molecular complexity.

Acknowledgments

The work described in this chapter was a collaborative effort by multiple investigators: Specifically, I'd like to thank Erika von Mutius and Ram Sasisekharan for their contributions to this work. This work was funded in part by the National Institutes of Health (T32 ES007020-38).

Chapter 10 : Thesis Summary

Glycans are an important class of biological molecules which regulate a variety of physiological processes ranging from signal transduction to tissue development and microbial pathogenesis. However, in comparison to DNA or proteins, elucidating glycan structure–function relationships presents unique challenges due to the structural complexity of glycans, the dominant role of multivalency in their sequence-specific interaction with glycan binding proteins (GBPs), and their “analog” modulation of biological function. To address these challenges, a unique approach which leverages the integration of a structural analysis of glycans, and glycan-proteins interactions with functional analysis across multiple level (genetic, cellular and organismal) is necessary to decode structure–function relationships. In this thesis, I develop new tools and implement integrated approaches to study glycans and glycan-binding proteins. Furthermore, I leverage these approaches to uncover new biological roles of glycans and GBPs in disease, specifically, focusing on the function of hemagglutinin-glycan receptor specificity in influenza pathogenesis and the function that glycans play in regulating cell-microenvironment interactions during cancer progression.

Section one of this thesis is focused on the development and application of tools to study glycan binding proteins. Glycan binding proteins, through their direct interaction with glycans, decode the information encoded within the glycome. In order to decode glycan structure-function relationships it is imperative to understand the structural basis for affinity and specificity of GBP-glycan interactions. To elucidate the structural basis for GBP-glycan interactions from the perspective of the glycan binding protein, it is important to characterize key structural attributes including: i) the glycan binding site, ii) key molecular contacts made in the glycan-protein interface, iii) key functional residues facilitating the GBP-glycan interaction, and iv) protein surface topology. Unfortunately, there are major limitations in characterizing these attributes, as current experimental methods are often low-throughput, technically challenging, or expensive and time consuming. Computational tools to aid in structural analysis of GBPs might prove useful in addressing these challenges by providing complementary information thereby reducing the amount of experimentations. Here, I implemented a computational tool that enabled protein inter-residue network analysis. I incorporated it alongside structural, biochemical and functional analysis to demonstrate the facile identification of key functional residues in the glycan binding site of a model glycan binding protein, FGF-2 (Chapter 2). These results suggest that SIN could be applied to the study of other glycan-binding proteins to elucidate structural determinants that are critical for glycan-GBP interactions.

Next, I applied this integrated approach to study the Influenza A Virus (IAV) hemagglutinin (HA) and the structural requirements for it to switch its receptor binding preferences from avian to human glycan receptors. Previously, our lab discovered that HAs recognize glycans based on their topology, where avian-adapted HAs recognize glycans with a cone-like topology and human-adapted HAs recognize glycans with an umbrella-like topology. Importantly, structural analysis revealed that HA-glycan complexes with umbrella-like glycans make different molecular interactions with amino acids in the RBS as compared to HA-glycan complexes with cone-like glycans. Furthermore, our lab developed a biochemical assay capable of assessing HA's affinity to cone or umbrella like glycans. From this, our lab found that human adaptation and airborne transmissibility was correlated with an HAs' quantitative switch in receptor specificity from cone- to umbrella-like glycans. With a more robust receptor definition, a biochemical assay to determine HA-glycan affinity, and a framework to correlate our biochemical readout to properties of host tropism and transmissibility, we asked an important question: What are the mutations required for that an avian-adapted HA to undergo a quantitative switch in affinity from avian to human receptors?

Towards addressing the above question, I reasoned that network analysis of HA would be useful to answer this question, particularly in conjunction with the tools above (i.e. insights into receptor topology & biochemical tools). In Chapter 2, I showed the inter-residue interaction network was able to identify key evolutionarily conserved regions (i.e. amino acid networks) of protein function in FGF-2 as well as the entire FGF family. Implicit in this notion is the idea that network analysis of protein structure can detect amino acid networks which are evolutionarily conserved in structure space, but not in a sequence space. Given the aforementioned, inter-residue network analysis could identify residues that are also structurally conserved (i.e. structurally constrained to mutate) in influenza HA. Additionally, network analysis allows for the mapping of inter-residue interaction onto a 2D network graph; this allow us to investigate qualitative network properties between human and avian adapted HA, which could prove useful in understanding the structural implication of particular HA mutations. To uncover the structural determinants of the quantitative switch we developed an approach which integrates multiple analyses, including: i) molecular modeling of HA-glycan complexes with cone-like or umbrella-like topologies to define structural features characteristic of human versus avian adapted HAs, ii) inter-residue network analysis to identify RBS positions that are constrained to mutate and assess RBS network graphs to define likely mutational paths that could give rise to structural features of human adaption, and iii) biochemical tools to create HA mutants that enable a quantitative switch in receptor binding specificity from avian to human glycan receptors.

With this integrated approach, we first studied the HAs from two influenza A viruses which pose a significant pandemic threat: H5N1 (bird flu) and the 2013 outbreak H7N9. For H5N1, we identified that four structural features necessary for human adaptation of H5N1. Next, by mapping mutational paths by which H5N1 could acquire these features and investigating the acquisition of these features in currently evolving H5N1 strains, we identified certain H5N1s have acquired one more of the structural features required for human adaptation. Finally, we show that in certain H5s, a single base pair mutation can quantitatively switch their binding to human receptors. Using a similar approach, we studied an HA from the 2013 outbreak H7N9 strain. Here, we report that while the H7N9 HA shows limited binding to human receptors, a single amino acid mutation would result in structural changes within the receptor binding site that allow for extensive binding to human receptors present in the upper respiratory tract.

Together these studies show our integrated approach to study HA-glycan receptor specificity provides a framework for improved pandemic surveillance, and ultimately could help identify IAV strains with high pandemic potential circulating in non-human hosts.

In addition to HA's role in host tropism, HA is a primary site for host immune recognition. Antigenic novelty is also a critical factor in viral pandemics. Thus, a robust pandemic surveillance framework should capture the antigenic properties alongside HA-glycan receptor specificity measurements. Towards this goal, we developed bioinformatics tools and implemented experimental methods capable of measuring antigenic properties into the pandemic risk assessment. Using this approach, we identified several H3 strains circulating in swine and birds that possessed hallmark features of viruses that could re-emerge and give rise to another H3N2 pandemic in humans (i.e. a high degree of antigenic similarity to the 1968 pandemic H3 and human glycan receptor binding).

Finally, leveraging the above tools and framework, we characterize the glycan receptor binding specificity and H3 from the 2011 seal H3N8, which caused a mass seal death event off the coast of New England. Initially this virus was concerning due to identification of mutations in the polymerase gene (PB2) that were associated with human adaptation. However, we report that seal H3N8 HA preferentially recognizes $\alpha 2\rightarrow 3$-linked glycans and thus is not likely infect humans.

In the work presented here, we effectively demonstrate a suite of complementary tools enabling us to study i) HA glycan binding specificity both biochemically & physiologically, ii) the structural determinants of HA-receptor specificity, iii) HA's mutational path to switch its receptor binding specificity, and iv) HA's antigenic properties. Taken together, our approach to studying HA-glycan interactions provides the foundation for improved pandemic surveillance and the identification of currently circulating influenza strains. Finally, applicability of the tools developed here is not restricted to influenza. These tools could ostensibly be applied to investigate other glycan-GBP interactions.

Section two of this focuses on the development and application of novel approaches to study of glycans and their biological function. To elucidate the functional role of glycans, a key first step is to measure their structural attributes, including: glycan fine-structure or "sequence" and the ensemble of glycans expressed (i.e. glycome). Due to glycan complexity there is no single tool capable of capturing all these features and multiple analytical tools and methods are needed. Here, I develop an integrated approach to characterize the cell surface glycome, including N-linked, O-linked glycans, and HSGAGs. This approach integrates glycogene expression data, analytical tools, and glycan motif binding protein reagents. Using this platform, I first demonstrate that rapid and efficient characterization the N- and O-linked glycome in a model cell system, representing metastatic and non-metastatic cells. The approach developed here identified changes to the glycome at multiple levels (i.e. changes in abundance of specific glycan and changes to glycan motifs), and could correlate these changes to the glycan biosynthesis genes responsible.

Next, I applied this this integrated approach to uncover a new role of glycans. Here, I study the role of HSGAGs in regulating cancer stem cell activity in breast cancer. We report that SULF1, an HSGAG-modifying enzyme, is required for efficient tumor initiation, growth and metastasis of CSCs. We further report that knockdown of SULF1 results in a mesenchymal-to-epithelial transition in CSCs via an increase of 6-O-sulfated HSGAGs which negatively regulate WNT and TGFβ signaling. We report that SULF1 is highly upregulated in tumors from patients with invasive breast carcinoma. Moreover, patients with ER-/PR- basal breast cancer that highly express SULF1 have a poorer prognosis compared to low SULF1 expressing tumors. These results suggest that SULF1 might play a role in human tumor progression. Taken together, the identification of SULF1's essential role in CSC physiology and regulation of the EMT, suggests that targeting SULF1 may be a viable therapeutic strategy to reduce CSC activity.

Finally, I apply my experience in developing integrated analytical tools to study how components in farm dust could be used as a therapeutic for allergic-type disease and asthma. Over the last three decades, an interesting phenomenon has emerged demonstrating that exposure to certain farm environments can protect against asthma and atopy. New evidence suggests that the protective effect can be conferred through exposure to dust isolated from cowsheds, leading to interest in developing a therapeutic extract made from farm dust. One challenge with this approach is that developing therapeutics from complex mixtures, where the active ingredient (API) is poorly defined, poses a significant regulatory challenge, specifically concerning safety and consistent potency. Interestingly, approved therapeutics such as Heparin or Glatiramer acetate faced similar challenges (i.e. inasmuch that they too are biologic substance, complex mixture, poorly defined API). In these instances, integrated analytical tools played a crucial role in defining the "product" and ensuring quality, consistency, and potency. Throughout my previous work, I developed substantial experience in developing integrated analytical strategies. Here, I apply a similar framework and toolset to conceptualize an integrated analytical strategy to decode composition-activity relationships of a farm dust extract. I outline an analytical workflow to assess product consistency and potency for therapeutic farm dust extracts with the aim of providing a regulatory path, enabling the development of farm dust extract-derived therapeutics.

Overall, this thesis provides important tools, approaches and insights to enable and improve the study of glycans and glycan binding proteins. Together, the work here provides a framework for decoding structure function relationship of glycans.

ADCC	Antibody dependent cell-mediated cytotoxicity
AHR	Airway hyper-responsiveness
AI	Antigenic identity
AIP-2	Anti-inflammatory protein-2
APC	Antigen presenting cell
API	Active Pharmaceutical Ingredient
ATR-FTIR	Attenuated total reflectance-Fourier transform infrared
bFGF	Basic fibroblast growth factor
BSA	Bovine serum albumin
CDC	Complement dependent cytotoxicity
CE	Capillary electrophoresis
CSC	cancer stem cell
DEF	Duck embryo fibroblasts
ECM	Extracellular matrix
EMT	Epithelial–mesenchymal transition
EPS	Extracellular polysaccharides
ER	Estrogen receptor
FACS	Fluorescence-activated cell sorting
FGFR	Fibroblast growth factor/ receptor
GA	Glatiramer acetate
GAG	Glycosaminoglycan
GBP	Glycan binding protien
GFP	Green flourescent protein
HA	Hemagglutinin
HDM	House Dust Mites
HLA-DR	Human Leukocyte Antigen - antigen D related
HMECs	Human mammary epithelia cells
HS/HSGAG	Heparin/heparan sulfate-like glycosaminoglycan
HSPG	Heparan sulfate proteoglycan
hTERT	Human telomerase
IAV	Influenza virus A
IEX HPLC	Ion-exchange high performance liquid chromatography
IFS	Inactivated Fetal Calf Serum
ILT	Immunoglobulin-like transcript
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPS	Lipopolysaccharide endotoxin
M1/2	Matrix protein 1/2
MALDI-TOF	Matrix-assisted laser desorption/ionization- time-of-Flight
MDCK	Madin-Darby Canine Kidney
MOI	Multiplicity of Infection
MS	Mass spectrometry

MW	Molecular weight
NA	Neuraminidase
Nd:YAG	Neodymium-doped yttrium aluminum garnet
NEP	Nuclear export protein
Neu5Ac	N-Acetylneuraminic acid
NMR	Nuclear magnetic resonance spectroscopy
NSCLC	Non-small cell lung cancer
OVA	Ovalbumin
PAMP	Pathogen-associated molecular patterns
PBL	Peripheral blood leukocyte
PDB	Protein Data Bank
PR	Progesterone receptor
PRR	Pattern recognition receptors
PTM	Post-translational modification
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
RBS	Receptor binding site
RNP	Viral Ribonucleoprotein
SEC	Size exclusion Chromatography
SIN	Significant Inter-residue Interaction Network
TCM	Traditional Chinese medicines
TLR	Toll-like receptors
TNBC	Triple negative breast cancers

References

[1] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, a G. Clark, J. Nadeau, V. a McKusick, N. Zinder, a J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab,

K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, a E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K. a Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V Merkulov, N. Milshina, H.M. Moore, a K. Naik, V. a Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, a Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, a Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, a Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigó, M.J. Campbell, K. V Sjolander, B. Karlak, a Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, a Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, a Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, a Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A.

Zandieh, X. Zhu, The sequence of the human genome., Science. 291 (2001) 1304–1351. doi:10.1126/science.1058040.

[2] A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler, eds., Essentials of Glycobiology, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2009.

[3] K.W. Moremen, M. Tiemeyer, A. V Nairn, Vertebrate protein glycosylation: diversity, synthesis and function., Nat. Rev. Mol. Cell Biol. 13 (2012) 448–462. doi:10.1038/nrm3383.

[4] R. Sasisekharan, R. Raman, V. Prabhakar, Glycomics approach to structure-function relationships of glycosaminoglycans., Annu. Rev. Biomed. Eng. 8 (2006) 181–231. doi:10.1146/annurev.bioeng.8.061505.095745.

[5] R. Sasisekharan, Z. Shriver, G. Venkataraman, U. Narayanasami, Roles of heparan-sulphate glycosaminoglycans in cancer., Nat. Rev. Cancer. 2 (2002) 521–528. doi:10.1038/nrc842.

[6] M. Pérez-Garay, B. Arteta, L. Pagès, R. de Llorens, C. de Bolòs, F. Vidal-Vanaclocha, R. Peracaula, alpha2,3-sialyltransferase ST3Gal III modulates pancreatic cancer cell motility and adhesion in vitro and enhances its metastatic potential in vivo., PLoS One. 5 (2010) e12524. doi:10.1371/journal.pone.0012524.

[7] Y. Dai, Y. Yang, V. MacLeod, X. Yue, A.C. Rapraeger, Z. Shriver, G. Venkataraman, R. Sasisekharan, R.D. Sanderson, HSulf-1 and HSulf-2 are potent inhibitors of myeloma tumor growth in vivo., J. Biol. Chem. 280 (2005) 40066–40073. doi:10.1074/jbc.M508136200.

[8] T. Hennet, Diseases of glycosylation beyond classical congenital disorders of glycosylation., Biochim. Biophys. Acta. 1820 (2012) 1306–1317. doi:10.1016/j.bbagen.2012.02.001.

[9] C.R. Holst, H. Bou-Reslan, B.B. Gore, K. Wong, D. Grant, S. Chalasani, R. a Carano, G.D. Frantz, M. Tessier-Lavigne, B. Bolon, D.M. French, A. Ashkenazi, Secreted sulfatases Sulf1 and Sulf2 have overlapping yet essential roles in mouse neonatal survival., PLoS One. 2 (2007) e575. doi:10.1371/journal.pone.0000575.

[10] M. Inatani, F. Irie, A.S. Plump, M. Tessier-Lavigne, Y. Yamaguchi, Mammalian brain morphogenesis and midline axon guidance require heparan sulfate., Science. 302 (2003) 1044–1046. doi:10.1126/science.1090497.

[11] R.D. Sanderson, Heparan sulfate proteoglycans in invasion and metastasis., Semin. Cell Dev. Biol. 12 (2001) 89–98. doi:10.1006/scdb.2000.0241.

[12] Y.-Y. Zhao, M. Takahashi, J.-G. Gu, E. Miyoshi, A. Matsumoto, S. Kitazume, N. Taniguchi, Functional roles of N-glycans in cell signaling and cell adhesion in cancer., Cancer Sci. 99 (2008) 1304–1310. doi:10.1111/j.1349-7006.2008.00839.x.

[13] N. Haines, K.D. Irvine, Glycosylation regulates Notch signalling., Nat. Rev. Mol. Cell Biol. 4 (2003) 786–797. doi:10.1038/nrm1228.

[14] D. Kolarich, B. Lepenies, P.H. Seeberger, Glycomics, glycoproteomics and the immune system., Curr. Opin. Chem. Biol. 16 (2012) 214–220. doi:10.1016/j.cbpa.2011.12.006.

[15] A. Chandrasekaran, A. Srinivasan, R. Raman, K. Viswanathan, S. Raguram, T.M. Tumpey, V. Sasisekharan, R. Sasisekharan, Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin., Nat. Biotechnol. 26 (2008) 107–113. doi:10.1038/nbt1375.

[16] A. Srinivasan, K. Viswanathan, R. Raman, A. Chandrasekaran, S. Raguram, T.M. Tumpey, V. Sasisekharan, R. Sasisekharan, Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 2800–2805. doi:10.1073/pnas.0711963105.

[17] A. V Nairn, K. Aoki, M. dela Rosa, M. Porterfield, J.-M. Lim, M. Kulik, J.M. Pierce, L. Wells, S. Dalton, M. Tiemeyer, K.W. Moremen, Regulation of glycan structures in murine embryonic stem cells: combined transcript profiling of glycan-related genes and glycan structural analysis., J. Biol. Chem. 287 (2012) 37835–37856. doi:10.1074/jbc.M112.405233.

[18] A. López-Ferrer, C. Barranco, C. de Bolós, Differences in the O-glycosylation patterns between lung squamous cell carcinoma and adenocarcinoma., Am. J. Clin. Pathol. 118 (2002) 749–755. doi:10.1309/LWP3-MFA8-8KX7-60YQ.

[19] S. Rosen, H. Lemjabbar-Alaoui, SULF-2: An extracellular modulator of cell signaling and a cancer target candidate, Expert Opin. Ther. Targets. 14 (2010) 935–949. doi:10.1517/14728222.2010.504718.SULF-2.

[20] T. Satomaa, A. Heiskanen, I. Leonardsson, J. Angström, A. Olonen, M. Blomqvist, N. Salovuori, C. Haglund, S. Teneberg, J. Natunen, O. Carpén, J. Saarinen, Analysis of the human cancer glycome identifies a novel group of tumor-associated N-acetylglucosamine glycan antigens., Cancer Res. 69 (2009) 5811–5819. doi:10.1158/0008-5472.CAN-08-0289.

[21] J. Goetz, Y. Mechref, P. Kang, M.-H. Jeng, M. V Novotny, Glycomic profiling of invasive and non-invasive breast cancer cells., Glycoconj. J. 26 (2009) 117–131. doi:10.1007/s10719-008-9170-4.

[22] J.N. Arnold, R. Saldova, U.M.A. Hamid, P.M. Rudd, Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation., Proteomics. 8 (2008) 3284–3293. doi:10.1002/pmic.200800163.

[23] D.J. Vigerust, V.L. Shepherd, Virus glycosylation: role in virulence and immune interactions., Trends Microbiol. 15 (2007) 211–218. doi:10.1016/j.tim.2007.03.003.

[24] D. Ghaderi, R.E.R. Taylor, V. Padler-Karavani, S. Diaz, A. Varki, Implications of the presence of N-glycolylneuraminic Acid in Recombinant Therapeutic Glycoproteins, Nat. Biotechnol. 28 (2010) 863–867. doi:10.1038/nbt.1651.Implications.

[25] S.J. Shire, W. Gombotz, K. Bechtold-Peters, J. Andya, eds., Current Trends in Monoclonal Antibody Development and Manufacturing, Springer New York, New York (NY), 2010. doi:10.1007/978-0-387-76643-0.

[26] R. Jefferis, Glycosylation as a strategy to improve antibody-based therapeutics., Nat. Rev. Drug Discov. 8 (2009) 226–234. doi:10.1038/nrd2804.

[27] J.N. Arnold, M.R. Wormald, R.B. Sim, P.M. Rudd, R.A. Dwek, The impact of glycosylation on the biological function and structure of human immunoglobulins., Annu. Rev. Immunol. 25 (2007) 21–50. doi:10.1146/annurev.immunol.25.022106.141702.

[28] R.A. Laine, Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10 12 structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems, Glycobiology. 4 (1994) 759–767. doi:10.1093/glycob/4.6.759.

[29] K. Mariño, J. Bones, J.J. Kattla, P.M. Rudd, A systematic approach to protein glycosylation analysis: a path through the maze., Nat. Chem. Biol. 6 (2010) 713–723. doi:10.1038/nchembio.437.

[30] R. Raman, S. Raguram, G. Venkataraman, J.C. Paulson, R. Sasisekharan, Glycomics : an integrated systems approach to structure-function relationships of glycans, Nat. Methods. 2 (2005) 817–824. doi:10.1038/NMETH807.

[31] M. Ambrosi, N.R. Cameron, B.G. Davis, Lectins: tools for the molecular understanding of the glycocode., Org. Biomol. Chem. 3 (2005) 1593–1608. doi:10.1039/b414350g.

[32] L.N. Robinson, C. Artpradit, R. Raman, Z.H. Shriver, M. Ruchirawat, R. Sasisekharan, Harnessing glycomics technologies: integrating structure with function for glycan characterization., Electrophoresis. 33 (2012) 797–814. doi:10.1002/elps.201100231.

[33] B.A. Cunha, Influenza: historical aspects of epidemics and pandemics, Infect. Dis. Clin. North Am. 18 (2004) 141–155. doi:10.1016/S0891-5520(03)00095-3.

[34] E.D. Kilbourne, Influenza pandemics of the 20th century, Emerg. Infect. Dis. 12 (2006) 9–14. doi:10.3201/eid1201.051254.

[35] J.K. Taubenberger, D.M. Morens, Influenza: the once and future pandemic., Public Health Rep. 125 Suppl (2010) 16–26.

[36] Y. Shi, Y. Wu, W. Zhang, J. Qi, G.F. Gao, Enabling the "host jump": structural determinants of receptor-binding specificity in influenza A viruses., Nat. Rev. Microbiol. 12 (2014) 822–31. doi:10.1038/nrmicro3362.

[37] T.M. Tumpey, T.R. Maines, N. Van Hoeven, L. Glaser, A. Solórzano, C. Pappas, N.J. Cox, D.E. Swayne, P. Palese, J.M. Katz, A. García-Sastre, A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission.,

Science. 315 (2007) 655–659. doi:10.1126/science.1136212.

[38] K. Tharakaraman, R. Raman, K. Viswanathan, N.W. Stebbins, A. Jayaraman, A. Krishnan, V. Sasisekharan, R. Sasisekharan, Structural Determinants for Naturally Evolving H5N1 Hemagglutinin to Switch Its Receptor Specificity, Cell. 153 (2013) 1475–1485. doi:10.1016/j.cell.2013.05.035.

[39] O. Hungnes, The role of genetic analysis in influenza virus surveillance and strain characterisation, Vaccine. 20 (2002) 45–49. doi:10.1016/S0264-410X(02)00515-7.

[40] A. Jayaraman, Engineering and Targeting Glycan Receptor Binding of Influenza A Virus Hemagglutinin, (2011).

[41] A. Srinivasan, K. Viswanathan, R. Raman, A. Chandrasekaran, S. Raguram, T.M. Tumpey, V. Sasisekharan, R. Sasisekharan, Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 2800–5. doi:10.1073/pnas.0711963105.

[42] M.I. Nelson, E.C. Holmes, The evolution of epidemic influenza., Nat. Rev. Genet. 8 (2007) 196–205. doi:10.1038/nrg2053.

[43] K. Tharakaraman, R. Raman, N.W. Stebbins, K. Viswanathan, V. Sasisekharan, R. Sasisekharan, Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic., Sci. Rep. 3 (2013) 1822. doi:10.1038/srep01822.

[44] R.A.M. Fouchier, P.M. Schneeberger, F.W. Rozendaal, J.M. Broekman, S.A.G. Kemink, V. Munster, T. Kuiken, G.F. Rimmelzwaan, M. Schutten, G.J.J. Van Doornum, G. Koch, A. Bosman, M. Koopmans, A.D.M.E. Osterhaus, Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome., Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 1356–61. doi:10.1073/pnas.0308352100.

[45] Y. Huang, X. Li, H. Zhang, B. Chen, Y. Jiang, L. Yang, W. Zhu, S. Hu, S. Zhou, Y. Tang, X. Xiang, F. Li, W. Li, L. Gao, Human infection with an avian influenza A (H9N2) virus in the middle region of China, J. Med. Virol. 87 (2015) 1641–1648. doi:10.1002/jmv.24231.

[46] S. Lai, Y. Qin, B.J. Cowling, X. Ren, N.A. Wardrop, M. Gilbert, T.K. Tsang, P. Wu, L. Feng, H. Jiang, Z. Peng, J. Zheng, Q. Liao, S. Li, P.W. Horby, J.J. Farrar, G.F. Gao, A.J. Tatem, H. Yu, Global epidemiology of avian influenza A H5N1 virus infection in humans, 1997–2015: a systematic review of individual case data, Lancet Infect. Dis. 16 (2016) e108–e118. doi:10.1016/S1473-3099(16)00153-5.

[47] Y.A. Shtyrya, L. V Mochalova, N. V Bovin, Influenza virus neuraminidase: structure and function., Acta Naturae. 1 (2009) 26–32. http://www.ncbi.nlm.nih.gov/pubmed/22649600%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3347517.

[48] W. Zheng, Y.J. Tao, Structure and assembly of the influenza A virus ribonucleoprotein complex, FEBS Lett. 587 (2013) 1206–1214. doi:10.1016/j.febslet.2013.02.048.

[49] M. Lakadamyali, M.J. Rust, X. Zhuang, Endocytosis of influenza viruses, Microbes Infect. 6 (2004) 929–936. doi:10.1016/j.micinf.2004.05.002.

[50] R.G. Webster, W. Bean, O. Gorman, T. Chambers, Y. Kawaoka, Evolution and Ecology of Influenza A Viruses, Am. Soc. Microbiol. 56 (1992) 152–179. doi:0146-0749/92/010152-28$02.00/0.

[51] J. Wahlgren, Influenza A viruses: an ecology review., Infect. Ecol. Epidemiol. 1 (2011) 1–7. doi:10.3402/iee.v1i0.6004.

[52] G.J.D. Smith, D. Vijaykrishna, J. Bahl, S.J. Lycett, M. Worobey, O.G. Pybus, S.K. Ma, C.L. Cheung, J. Raghwani, S. Bhatt, J.S.M. Peiris, Y. Guan, A. Rambaut, Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic., Nature. 459 (2009) 1122–1125. doi:10.1038/nature08182.

[53] K. Nakajima, U. Desselberger, P. Palese, Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950., Nature. 274 (1978) 334–339. doi:10.1038/274334a0.

[54] D.J. Smith, Mapping the Antigenic and Genetic Evolution of Influenza Virus, Science (80-.). 305 (2004) 371–376. doi:10.1126/science.1097211.

[55] J. Stech, X. Xiong, C. Scholtissek, R.G. Webster, Independence of evolutionary and mutational rates after transmission of avian influenza viruses to swine, J Virol.

73 (1999) 1878–1884. http://www.ncbi.nlm.nih.gov/pubmed/9971766.

[56] F. CARRAT, A. FLAHAULT, Influenza vaccine: The challenge of antigenic drift, Vaccine. 25 (2007) 6852–6862. doi:10.1016/j.vaccine.2007.07.027.

[57] M. de Graaf, R.A.M. Fouchier, Role of receptor binding specificity in influenza A virus transmission and pathogenesis, Embo J. 33 (2014) 823–841. doi:10.1002/embj.201387442.

[58] R. Raman, K. Tharakaraman, Z. Shriver, A. Jayaraman, V. Sasisekharan, R. Sasisekharan, Glycan receptor specificity as a useful tool for characterization and surveillance of influenza A virus, Trends Microbiol. 22 (2014) 632–641. doi:10.1016/j.tim.2014.07.002.

[59] I.A. WILSON, J.J. SKEHEL, D.C. WILEY, Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 AA resolution., Nature. 289 (1981) 366-- 73. http://caslon.stanford.edu:3210/sfxlcl3?url_ver=Z39.88-2004&ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.issn=0028-0836&rft.coden=NATUAS&rft.date=1981&rft.volume=289&rft.issue=5796&rft.spage=366&rft.epage=73&rft.atitle=Structure+of+the+haemagglutinin+membrane+glycoprotein+of+influenza+%25Avirus+at+3+AA+resolution%252E&rft.jtitle=Nature&rft.aulast=WILSON&rft.auinit=IA&rft.pub=UK+%253A+1981&rft_id=info:lanl-repo/inspec/1675219&rft_id=info:

[60] E. Hoffmann, G. Neumann, Y. Kawaoka, G. Hobom, R.G. Webster, A DNA transfection system for generation of influenza A virus from eight plasmids, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 6108–6113. doi:10.1073/pnas.100133697.

[61] S. Pleschka, R. Jaskunas, O.G. Engelhardt, T. Zürcher, P. Palese, A. García-Sastre, A plasmid-based reverse genetics system for influenza A virus., J. Virol. 70 (1996) 4188–92. http://www.ncbi.nlm.nih.gov/pubmed/8648766%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC190316.

[62] C. Pappas, P. V Aguilar, C.F. Basler, A. Solórzano, H. Zeng, L.A. Perrone, P. Palese, A. García-Sastre, J.M. Katz, T.M. Tumpey, Single gene reassortants

identify a critical role for PB1, HA, and NA in the high virulence of the 1918 pandemic influenza virus., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 3064–9. doi:10.1073/pnas.0711815105.

[63] T.M. Tumpey, A. Garcia-Sastre, J.K. Taubenberger, P. Palese, D.E. Swayne, C.F. Basler, Pathogenicity and immunogenicity of influenza viruses with genes from the 1918 pandemic virus, Proc Natl Acad Sci U S A. 101 (2004) 3166–3171. doi:10.1073/pnas.0308391100.

[64] P.B. Gupta, C.L. Chaffer, R.A. Weinberg, Cancer stem cells: mirage or reality?, Nat. Med. 15 (2009) 1010–2. doi:10.1038/nm0909-1010.

[65] J.E. Visvader, G.J. Lindeman, Cancer stem cells in solid tumours: accumulating evidence and unresolved questions., Nat. Rev. Cancer. 8 (2008) 755–68. doi:10.1038/nrc2499.

[66] M. Al-Hajj, M.S. Wicha, A. Benito-Hernandez, S.J. Morrison, M.F. Clarke, Prospective identification of tumorigenic breast cancer cells, Proc. .... 100 (2003). http://www.pnas.org/content/100/7/3983.short (accessed September 2, 2013).

[67] M. Kai, N. Kanaya, S. V. Wu, C. Mendez, D. Nguyen, T. Luu, S. Chen, Targeting breast cancer stem cells in triple-negative breast cancer using a combination of LBH589 and salinomycin, Breast Cancer Res. Treat. 151 (2015) 281–294. doi:10.1007/s10549-015-3376-5.

[68] V.P.S. Ghotra, J.C. Puigvert, E.H.J. Danen, The cancer stem cell microenvironment and anti-cancer therapy., Int. J. Radiat. Biol. 85 (2009) 955–962. doi:10.3109/09553000903242164.

[69] C. Scheel, E.N. Eaton, S.H.-J. Li, C.L. Chaffer, F. Reinhardt, K.-J. Kah, G. Bell, W. Guo, J. Rubin, A.L. Richardson, R.A. Weinberg, Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast., Cell. 145 (2011) 926–940. doi:10.1016/j.cell.2011.04.029.

[70] H. Lu, K.R. Clauser, W.L. Tam, J. Fröse, X. Ye, E.N. Eaton, F. Reinhardt, V.S. Donnenberg, R. Bhargava, S.A. Carr, R.A. Weinberg, A breast cancer stem cell niche supported by juxtacrine signalling from monocytes and macrophages., Nat. Cell Biol. 16 (2014) 1105–17. doi:10.1038/ncb3041.

[71] N. Kannan, L. V. Nguyen, C.J. Eaves, Integrin β3 links therapy resistance and cancer stem cell properties, Nat. Cell Biol. 16 (2014) 397–399. doi:10.1038/ncb2960.

[72] G.B. Adams, D.T. Scadden, A niche opportunity for stem cell therapeutics., Gene Ther. 15 (2008) 96–99. doi:10.1038/sj.gt.3303063.

[73] D.L. Dragu, L.G. Necula, C. Bleotu, C.C. Diaconu, M. Chivu-Economescu, Therapies targeting cancer stem cells: Current trends and future challenges., World J. Stem Cells. 7 (2015) 1185–201. doi:10.4252/wjsc.v7.i9.1185.

[74] K. Kise, Y. Kinugasa-Katayama, N. Takakura, Tumor microenvironment for cancer stem cells, Adv. Drug Deliv. Rev. 99 (2016) 197–205. doi:10.1016/j.addr.2015.08.005.

[75] E. a Vasievich, L. Huang, The suppressive tumor microenvironment: a challenge in cancer immunotherapy., Mol. Pharm. 8 (2011) 635–41. doi:10.1021/mp1004228.

[76] N.W. Stebbins, R. Sasisekharan, Global Glycomics Analysis: Methodologies and Challenges, in: Fundam. Adv. Omi. Technol. From Genes to Metab., 2013.

[77] D.M. Beauvais, A.C. Rapraeger, Syndecans in tumor cell adhesion and signaling., Reprod. Biol. Endocrinol. 2 (2004) 3. doi:10.1186/1477-7827-2-3.

[78] P. Bashkin, S. Doctrow, M. Klagsbrun, C.M. Svahn, J. Folkman, I. Vlodavsky, Basic fibroblast growth factor binds to subendothelial extracellular matrix and is released by heparitinase and heparin-like molecules., Biochemistry. 28 (1989) 1737–1743. doi:10.1021/bi00430a047.

[79] T. Kiziltepe, Nitric Oxide-Induced DNA Recombination & Glycosaminosglycan Mediated Differentiation in Stem Cells, MIT, 2005.

[80] U. Barash, V. Cohen-Kaplan, I. Dowek, R.D. Sanderson, N. Ilan, I. Vlodavsky, V. Cohen-Kaplan, Proteoglycans in health and disease: new concepts for heparanase function in tumor progression and metastasis., FEBS J. 277 (2010) 3890–903. doi:10.1111/j.1742-4658.2010.07799.x.

[81] C.J. Jones, S. Beni, J.F.K. Limtiaco, D.J. Langeslay, C.K. Larive, Heparin characterization: challenges and solutions., Annu. Rev. Anal. Chem. (Palo Alto.

Calif). 4 (2011) 439–65. doi:10.1146/annurev-anchem-061010-113911.

[82] J.E. Turnbull, R. Sasisekharan, Glycomics: technologies taming a frontier omics field., OMICS. 14 (2010) 385–7. doi:10.1089/omi.2010.0067.

[83] J. Turnbull, a Powell, S. Guimond, Heparan sulfate: decoding a dynamic multifunctional cell regulator., Trends Cell Biol. 11 (2001) 75–82. http://www.ncbi.nlm.nih.gov/pubmed/11166215.

[84] M. Kusche-Gullberg, Sulfotransferases in glycosaminoglycan biosynthesis, Curr. Opin. Struct. Biol. 13 (2003) 605–611. doi:10.1016/j.sbi.2003.08.002.

[85] J.D. Esko, S.B. Selleck, Order out of chaos: assembly of ligand binding sites in heparan sulfate., Annu. Rev. Biochem. 71 (2002) 435–471. doi:10.1146/annurev.biochem.71.110601.135458.

[86] W.C. Lamanna, I. Kalus, M. Padva, R.J. Baldwin, C.L.R. Merry, T. Dierks, The heparanome--the enigma of encoding and decoding heparan sulfate sulfation., J. Biotechnol. 129 (2007) 290–307. doi:10.1016/j.jbiotec.2007.01.022.

[87] J. Schlessinger, a N. Plotnikov, O. a Ibrahimi, a V Eliseenkova, B.K. Yeh, a Yayon, R.J. Linhardt, M. Mohammadi, Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization., Mol. Cell. 6 (2000) 743–50. http://www.ncbi.nlm.nih.gov/pubmed/11030354.

[88] G.K. Dhoot, M.K. Gustafsson, X. Ai, W. Sun, D.M. Standiford, C.P. Emerson, Regulation of Wnt signaling and embryo patterning by an extracellular sulfatase., Science (80-.). 293 (2001) 1663–1666. doi:10.1126/science.293.5535.1663.

[89] X. Ai, A.-T. Do, O. Lozynska, M. Kusche-Gullberg, U. Lindahl, C.P. Emerson, QSulf1 remodels the 6-O sulfation states of cell surface heparan sulfate proteoglycans to promote Wnt signaling., J. Cell Biol. 162 (2003) 341–351. doi:10.1083/jcb.200212083.

[90] M. Petitou, B. Casu, U. Lindahl, 1976-1983, a critical period in the history of heparin: The discovery of the antithrombin binding site, Biochimie. 85 (2003) 83–89. doi:10.1016/S0300-9084(03)00078-6.

[91] S. Guglier, M. Hricovíni, R. Raman, L. Polito, G. Torri, B. Casu, R. Sasisekharan,

M. Guerrini, Minimum FGF2 binding structural requirements of heparin and heparan sulfate oligosaccharides as determined by NMR spectroscopy., Biochemistry. 47 (2008) 13862–13869. http://www.ncbi.nlm.nih.gov/pubmed/19117094.

[92] R. Raman, G. Venkataraman, S. Ernst, V. Sasisekharan, R. Sasisekharan, Structural specificity of heparin binding in the fibroblast growth factor family of proteins., Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 2357–2362. doi:10.1073/pnas.0437842100.

[93] D. Liu, Z. Shriver, G. Venkataraman, Y. El Shabrawi, R. Sasisekharan, Tumor cell surface heparan sulfate as cryptic promoters or inhibitors of tumor growth and metastasis., Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 568–73. doi:10.1073/pnas.012578299.

[94] M. Morimoto-Tomita, K. Uchimura, Z. Werb, S. Hemmerich, S.D. Rosen, Cloning and characterization of two extracellular heparin-degrading endosulfatases in mice and humans., J. Biol. Chem. 277 (2002) 49175–49185. doi:10.1074/jbc.M205131200.

[95] J.-P. Lai, S.S. Dalbir, S. Abdirashid M., L.R. Roberts, The Tumor Suppressor Function of Human Sulfatase 1 (SULF1) in Carcinogenesis, J Gastrointest Cancer. 39 (2008) 149–158. doi:10.1007/s12029-009-9058-y.

[96] C. Bret, J. Moreaux, J.-F. Schved, D. Hose, B. Klein, SULFs in human neoplasia: implication as progression and prognosis factors., J. Transl. Med. 9 (2011). doi:10.1186/1479-5876-9-72.

[97] L. Lundin, H. Larsson, J. Kreuger, S. Kanda, U. Lindahl, M. Salmivirta, L. Claesson-Welsh, Selectively desulfated heparin inhibits fibroblast growth factor-induced mitogenicity and angiogenesis., J. Biol. Chem. 275 (2000) 24653–24660. doi:10.1074/jbc.M908930199.

[98] H. Lemjabbar-alaoui, A. Van Zante, M.S. Singer, Q. Xue, Y.-Q. Wang, D. Tsay, B. He, D.M. Jablons, S.D. Rosen, Sulf-2, a heparan sulfate endosulfatase, promotes human lung carcinogenesis, Oncogene. 29 (2010) 635–646. doi:10.1038/onc.2009.365.

[99] I. Vlodavsky, Y. Friedmann, M. Elkin, H. Aingorn, R. Atzmon, R. Ishai-Michaeli, M. Bitan, O. Pappo, T. Peretz, I. Michal, L. Spector, I. Pecker, Mammalian heparanase: gene cloning, expression and function in tumor progression and metastasis., Nat. Med. 5 (1999) 793–802. doi:10.1038/10518.

[100] E. Cohen, I. Doweck, I. Naroditsky, O. Ben-Izhak, Heparanase is over-expressed in lung cancer and inversely correlates with patient's survival, Cancer. 113 (2008) 1004–1011. doi:10.1002/cncr.23680.Heparanase.

[101] A. Purushothaman, T. Uyama, F. Kobayashi, S. Yamada, K. Sugahara, A.C. Rapraeger, R.D. Sanderson, Heparanase-enhanced shedding of syndecan-1 by myeloma cells promotes endothelial invasion and angiogenesis., Blood. 115 (2010) 2449–57. doi:10.1182/blood-2009-07-234757.

[102] J.P. Ritchie, V.C. Ramani, Y. Ren, A. Naggi, G. Torri, B. Casu, S. Penco, C. Pisano, P. Carminati, M. Tortoreto, F. Zunino, I. Vlodavsky, R.D. Sanderson, Y. Yang, SST0001, a chemically modified heparin, inhibits myeloma growth and angiogenesis via disruption of the heparanase/syndecan-1 axis., Clin. Cancer Res. 17 (2011) 1382–93. doi:10.1158/1078-0432.CCR-10-2476.

[103] K. Dredge, E. Hammond, P. Handley, T.J. Gonda, M.T. Smith, C. Vincent, R. Brandt, V. Ferro, I. Bytheway, PG545, a dual heparanase and angiogenesis inhibitor, induces potent anti-tumour and anti-metastatic efficacy in preclinical models., Br. J. Cancer. 104 (2011) 635–42. doi:10.1038/bjc.2011.11.

[104] C.J. Liu, J. Chang, P.H. Lee, D.Y. Lin, C.C. Wu, L. Bin Jeng, Y.J. Lin, K.T. Mok, W.C. Lee, H.Z. Yeh, M.C. Ho, S.S. Yang, M.D. Yang, M.C. Yu, R.H. Hu, C.Y. Peng, K.L. Lai, S.S.C. Chang, P.J. Chen, Adjuvant heparanase inhibitor PI-88 therapy for hepatocellular carcinoma recurrence, World J. Gastroenterol. 20 (2014) 11381–11393. doi:10.3748/wjg.v20.i32.11384.

[105] R.J. Linhardt, J.E. Turnbull, H.M. Wang, D. Loganathan, J.T. Gallagher, Examination of the substrate specificity of heparin and heparan sulfate lyases., Biochemistry. 29 (1990) 2611–2617. doi:10.1021/bi00462a026.

[106] G. Venkataraman, Z. Shriver, R. Raman, R. Sasisekharan, Sequencing complex polysaccharides., Science. 286 (1999) 537–42. c:%5CDocuments and

Settings%5CJongyoon Han%5CDesktop%5CPDF%5CNanofilter%5C1999 Science RAM saccharide sequencing.pdf.

[107] a J. Rhomberg, S. Ernst, R. Sasisekharan, K. Biemann, Mass spectrometric and capillary electrophoretic investigation of the enzymatic degradation of heparin-like glycosaminoglycans., Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 4176–4181. doi:10.1073/pnas.95.8.4176.

[108] V. Prabhakar, I. Capila, R. Sasisekharan, The Structural Elucidation of Glycosaminoglycans, in: N.H. Packer, N.G. Karlsson (Eds.), Methods Mol. Biol. Glycomics Methods Protoc., Humana Press, Totowa, NJ, 2009: pp. 147–156. doi:10.1007/978-1-59745-022-5.

[109] Y. Chang, B. Yang, X. Zhao, R.J. Linhardt, Analysis of glycosaminoglycan-derived disaccharides by capillary electrophoresis using laser-induced fluorescence detection., Anal. Biochem. 427 (2012) 91–8. doi:10.1016/j.ab.2012.05.004.

[110] R. Raman, V. Sasisekharan, R. Sasisekharan, Structural insights into biological roles of protein-glycosaminoglycan interactions., Chem. Biol. 12 (2005) 267–77. doi:10.1016/j.chembiol.2004.11.020.

[111] W.C. Lamanna, M.-A. Frese, M. Balleininger, T. Dierks, Sulf loss influences N-, 2-O-, and 6-O-sulfation of multiple heparan sulfate proteoglycans and modulates fibroblast growth factor signaling., J. Biol. Chem. 283 (2008) 27724–35. doi:10.1074/jbc.M802130200.

[112] D. Bonnet, J. Dick, Human Acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell, Nat. Med. 3 (1997) 730–737.

[113] R. Yamamoto, Y. Morita, J. Ooehara, S. Hamanaka, M. Onodera, K.L. Rudolph, H. Ema, H. Nakauchi, Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells., Cell. 154 (2013) 1112–26. doi:10.1016/j.cell.2013.08.007.

[114] J. Chen, Y. Li, T.-S. Yu, R.M. McKay, D.K. Burns, S.G. Kernie, L.F. Parada, A restricted cell population propagates glioblastoma growth after chemotherapy, Nature. 488 (2012) 522–526. doi:10.1038/nature11287.

[115] D.R. Pattabiraman, R.A. Weinberg, Tackling the cancer stem cells - what

challenges do they pose?, Nat Rev Drug Discov. 13 (2014) 497–512. doi:10.1038/nrd4253.

[116] S. a Mani, W. Guo, M.-J. Liao, E.N. Eaton, A. Ayyanan, A.Y. Zhou, M. Brooks, F. Reinhard, C.C. Zhang, M. Shipitsin, L.L. Campbell, K. Polyak, C. Brisken, J. Yang, R. a Weinberg, The epithelial-mesenchymal transition generates cells with properties of stem cells., Cell. 133 (2008) 704–15. doi:10.1016/j.cell.2008.03.027.

[117] K.M. Britton, J. a. Kirby, T.W.J. Lennard, A.P. Meeson, Cancer Stem Cells and Side Population Cells in Breast Cancer and Metastasis, Cancers (Basel). 3 (2011) 2106–2130. doi:10.3390/cancers3022106.

[118] B. De Craene, G. Berx, Regulatory networks defining EMT during cancer initiation and progression, Nat. Rev. Cancer. 13 (2013) 97–110. doi:10.1038/nrc3447.

[119] C.L. Chaffer, I. Brueckmann, C. Scheel, A.J. Kaestli, P.A. Wiggins, L.O. Rodrigues, M. Brooks, F. Reinhardt, Y. Su, K. Polyak, L.M. Arendt, C. Kuperwasser, B. Bierie, R. a Weinberg, Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state., Proc. Natl. Acad. Sci. 108 (2011) 7950–7955. doi:10.1073/pnas.1102454108.

[120] C.L. Chaffer, B.P. San Juan, E. Lim, R.A. Weinberg, EMT, cell plasticity and metastasis, Cancer Metastasis Rev. (2016) 1–10. doi:10.1007/s10555-016-9648-7.

[121] J. Zhang, X. Tian, J. Xing, Signal Transduction Pathways of EMT Induced by TGF-β, SHH, and WNT and Their Crosstalks, J. Clin. Med. 5 (2016) 41. doi:10.3390/jcm5040041.

[122] A. Beenken, M. Mohammadi, The FGF family: biology, pathophysiology and therapy., Nat. Rev. Drug Discov. 8 (2009) 235–53. doi:10.1038/nrd2792.

[123] R. PADERA, G. Venkataraman, D. Berry, R. Godavarti, R. Sasisekharan, FGF-2/fibroblast growth factor receptor/heparin-like glycosaminoglycan interactions: a compensation model for FGF-2 signaling, FASEB J. 13 (1999) 1677–1687. http://www.fasebj.org/content/13/13/1677 (accessed January 21, 2013).

[124] T. Puvirajesinghe, J. Turnbull, Glycoarray Technologies: Deciphering Interactions from Proteins to Live Cell Responses, Microarrays. 5 (2016) 3.

doi:10.3390/microarrays5010003.

[125] A. Ori, P. Free, J. Courty, M.C. Wilkinson, D.G. Fernig, Identification of heparin-binding sites in proteins by selective labeling., Mol. Cell. Proteomics. 8 (2009) 2256–2265. doi:10.1074/mcp.M900031-MCP200.

[126] L.H. Greene, Protein structure networks, Brief. Funct. Genomics. 11 (2012) 469–478. doi:10.1093/bfgp/els039.

[127] M.P. Cusack, B. Thibert, D.E. Bredesen, G. del Rio, Efficient identification of critical residues based only on protein structure by network analysis, PLoS One. 2 (2007) 1–7. doi:10.1371/journal.pone.0000421.

[128] G. Hu, W. Yan, J. Zhou, B. Shen, Residue interaction network analysis of Dronpa and a DNA clamp, J. Theor. Biol. 348 (2014) 55–64. doi:10.1016/j.jtbi.2014.01.023.

[129] V. Soundararajan, S. Zheng, N. Patel, K. Warnock, R. Raman, I.A. Wilson, S. Raguram, V. Sasisekharan, R. Sasisekharan, Networks link antigenic and receptor-binding sites of influenza hemagglutinin: mechanistic insight into fitter strain propagation., Sci. Rep. 1 (2011) 200. doi:10.1038/srep00200.

[130] S.E. Hensley, S.R. Das, A.L. Bailey, L.M. Schmidt, H.D. Hickman, A. Jayaraman, K. Viswanathan, R. Raman, R. Sasisekharan, J.R. Bennink, J.W. Yewdell, Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift., Science. 326 (2009) 734–6. doi:10.1126/science.1178258.

[131] R. Goetz, M. Mohammadi, Exploring mechanisms of FGF signalling through the lens of structural biology, Nat. Rev. Mol. Cell Biol. 14 (2013) 166–180. doi:10.1038/nrm3528.

[132] X. Zhang, O.A. Ibrahimi, S.K. Olsen, H. Umemori, M. Mohammadi, D.M. Ornitz, Receptor specificity of the fibroblast growth factor family: The complete mammalian FGF family, J. Biol. Chem. 281 (2006) 15694–15700. doi:10.1074/jbc.M601252200.

[133] G. Venkataraman, R. Raman, V. Sasisekharan, R. Sasisekharan, Molecular characteristics of fibroblast growth factor-fibroblast growth factor receptor-heparin-like glycosaminoglycan complex, Proc. Natl. Acad. Sci. 96 (1999) 3658–3663.

doi:10.1073/pnas.96.7.3658.

[134] E.M. Muñoz, R.J. Linhardt, Heparin-binding domains in vascular biology, Arterioscler. Thromb. Vasc. Biol. 24 (2004) 1549–1557. doi:10.1161/01.ATV.0000137189.22999.3f.

[135] Y. Guan, G.J.D. Smith, R. Webby, R.G. Webster, Molecular epidemiology of H5N1 avian influenza., Rev. Sci. Tech. - Off. Int. Des Épizooties. 28 (2009) 39–47. http://www.ncbi.nlm.nih.gov/pubmed/19618617 (accessed November 15, 2016).

[136] G. Neumann, H. Chen, G.F. Gao, Y. Shu, Y. Kawaoka, H5N1 influenza viruses: outbreaks and biological properties., Cell Res. 20 (2010) 51–61. doi:10.1038/cr.2009.124.

[137] S. Ge, Z. Wang, An overview of influenza A virus receptors., Crit. Rev. Microbiol. 37 (2011) 157–65. doi:10.3109/1040841X.2010.536523.

[138] C. Pappas, K. Viswanathan, A. Chandrasekaran, R. Raman, J.M. Katz, R. Sasisekharan, T.M. Tumpey, Receptor specificity and transmission of H2N2 subtype viruses isolated from the pandemic of 1957., PLoS One. 5 (2010) e11158. doi:10.1371/journal.pone.0011158.

[139] K. Viswanathan, X. Koh, A. Chandrasekaran, C. Pappas, R. Raman, A. Srinivasan, Z. Shriver, T.M. Tumpey, R. Sasisekharan, Determinants of glycan receptor specificity of H2N2 influenza A virus hemagglutinin., PLoS One. 5 (2010) e13768. doi:10.1371/journal.pone.0013768.

[140] Z. Shriver, R. Raman, K. Viswanathan, R. Sasisekharan, Context-specific target definition in influenza a virus hemagglutinin-glycan receptor interactions., Chem. Biol. 16 (2009) 803–14. doi:10.1016/j.chembiol.2009.08.002.

[141] S.J. Gamblin, L.F. Haire, R.J. Russell, D.J. Stevens, B. Xiao, Y. Ha, N. Vasisht, D.A. Steinhauer, R.S. Daniels, A. Elliot, D.C. Wiley, J.J. Skehel, The structure and receptor binding properties of the 1918 influenza hemagglutinin., Science. 303 (2004) 1838–1842. doi:10.1126/science.1093155.

[142] J. Liu, D.J. Stevens, L.F. Haire, P.A. Walker, P.J. Coombs, R.J. Russell, S.J. Gamblin, J.J. Skehel, Structures of receptor complexes formed by hemagglutinins

from the Asian Influenza pandemic of 1957., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 17175–80. doi:10.1073/pnas.0906849106.

[143] Y. Ha, D.J. Stevens, J.J. Skehel, D.C. Wiley, X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs., Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 11181–6. doi:10.1073/pnas.201401198.

[144] Y.P. Lin, X. Xiong, S.A. Wharton, S.R. Martin, P.J. Coombs, S.G. Vachieri, E. Christodoulou, P.A. Walker, J. Liu, J.J. Skehel, S.J. Gamblin, A.J. Hay, R.S. Daniels, J.W. McCauley, Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 21474–9. doi:10.1073/pnas.1218841110.

[145] A. Gambaryan, A. Tuzikov, G. Pazynina, N. Bovin, A. Balish, A. Klimov, Evolution of the receptor binding phenotype of influenza A (H5) viruses., Virology. 344 (2006) 432–8. doi:10.1016/j.virol.2005.08.035.

[146] J. Stevens, O. Blixt, T.M. Tumpey, J.K. Taubenberger, J.C. Paulson, I.A. Wilson, Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus., Science. 312 (2006) 404–410. doi:10.1126/science.1124513.

[147] J. Stevens, O. Blixt, L.-M. Chen, R.O. Donis, J.C. Paulson, I.A. Wilson, Recent avian H5N1 viruses exhibit increased propensity for acquiring human receptor specificity., J. Mol. Biol. 381 (2008) 1382–1394. doi:10.1016/j.jmb.2008.04.016.

[148] C.-C. Wang, J.-R. Chen, Y.-C. Tseng, C.-H. Hsu, Y.-F. Hung, S.-W. Chen, C.-M. Chen, K.-H. Khoo, T.-J. Cheng, Y.-S.E. Cheng, J.-T. Jan, C.-Y. Wu, C. Ma, C.-H. Wong, Glycans on influenza hemagglutinin affect receptor binding and immune response., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 18137–42. doi:10.1073/pnas.0909696106.

[149] Y. Watanabe, M.S. Ibrahim, Y. Suzuki, K. Ikuta, The changing nature of avian influenza A virus (H5N1)., Trends Microbiol. 20 (2012) 11–20. doi:10.1016/j.tim.2011.10.003.

[150] S. Yamada, Y. Suzuki, T. Suzuki, M.Q. Le, C.A. Nidom, Y. Sakai-Tagawa, Y. Muramoto, M. Ito, M.M. Kiso, T. Horimoto, K. Shinya, T. Sawada, M.M. Kiso, T.

Usui, T. Murata, Y. Lin, A. Hay, L.F. Haire, D.J. Stevens, R.J. Russell, S.J. Gamblin, J.J. Skehel, Y. Kawaoka, Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors., Nature. 444 (2006) 378–382. doi:10.1038/nature05264.

[151] M. Imai, T. Watanabe, M. Hatta, S.C. Das, M. Ozawa, K. Shinya, G. Zhong, A. Hanson, H. Katsura, S. Watanabe, C. Li, E. Kawakami, S. Yamada, M. Kiso, Y. Suzuki, E.A. Maher, G. Neumann, Y. Kawaoka, Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets., Nature. 486 (2012) 420–8. doi:10.1038/nature10831.

[152] S. Herfst, E.J. a Schrauwen, M. Linster, S. Chutinimitkul, E. de Wit, V.J. Munster, E.M. Sorrell, T.M. Bestebroer, D.F. Burke, D.J. Smith, G.F. Rimmelzwaan, A.D.M.E. Osterhaus, R. a M. Fouchier, Airborne transmission of influenza A/H5N1 virus between ferrets., Science. 336 (2012) 1534–41. doi:10.1126/science.1213362.

[153] Y. Watanabe, M.S. Ibrahim, H.F. Ellakany, N. Kawashita, R. Mizuike, H. Hiramatsu, N. Sriwilaijaroen, T. Takagi, Y. Suzuki, K. Ikuta, Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt., PLoS Pathog. 7 (2011) e1002068. doi:10.1371/journal.ppat.1002068.

[154] A. Jayaraman, C. Pappas, R. Raman, J.A. Belser, K. Viswanathan, Z. Shriver, T.M. Tumpey, R. Sasisekharan, A single base-pair change in 2009 H1N1 hemagglutinin increases human receptor affinity and leads to efficient airborne viral transmission in ferrets., PLoS One. 6 (2011) e17616. doi:10.1371/journal.pone.0017616.

[155] T.R. Maines, A. Jayaraman, J.A. Belser, D.A. Wadford, C. Pappas, H. Zeng, K.M. Gustin, M.B. Pearce, K. Viswanathan, Z.H. Shriver, R. Raman, N.J. Cox, R. Sasisekharan, J.M. Katz, T.M. Tumpey, Transmission and pathogenesis of swine-origin 2009 A(H1N1) influenza viruses in ferrets and mice., Science. 325 (2009) 484–7. doi:10.1126/science.1177238.

[156] M. Pearce, A. Jayaraman, Pathogenesis and transmission of swine origin A

(H3N2)v influenza viruses in ferrets, Proc. .... (2012). doi:10.1073/pnas.1119945109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1119945109.

[157] V. Soundararajan, S. Zheng, N. Patel, K. Warnock, R. Raman, I.A. Wilson, S. Raguram, V. Sasisekharan, R. Sasisekharan, J.J. Skehel, D.C. Wiley, J. Stevens, O. Blixt, T.M. Tumpey, J.K. Taubenberger, J.C. Paulson, I.A. Wilson, T.M. Tumpey, J.W. Yewdell, R.G. Webster, W.U. Gerhard, A.J. Caton, G.G. Brownlee, J.W. Yewdell, W. Gerhard, J.R. Gog, N.M. Ferguson, A.P. Galvani, R.M. Bush, D. Fleury, B. Barrère, T. Bizebard, R.S. Daniels, J.J. Skehel, M. Knossow, I.A. Wilson, J.J. Skehel, D.C. Wiley, S.E. Hensley, N. Tokuriki, C. Oldfield, E. Domingo, J.J. Holland, V. Soundararajan, D.J. Smith, C.J. Wei, S.R. Das, P. Puigbò, S.E. Hensley, D.E. Hurt, J.R. Bennink, J.W. Yewdell, P.D. Kwong, I.A. Wilson, V. Soundararajan, R. Raman, S. Raguram, V. Sasisekharan, R. Sasisekharan, N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, V. Soundararajan, N. Patel, V. Subramanian, V. Sasisekharan, R. Sasisekharan, U. Alon, M.B. Eisen, S. Sabesan, J.J. Skehel, D.C. Wiley, A. Jayaraman, J. Stevens, O. Blixt, J.C. Paulson, I.A. Wilson, R. Xu, D.C. Ekiert, J.C. Krause, R. Hai, J.E. Crowe, I.A. Wilson, Networks link antigenic and receptor-binding sites of influenza hemagglutinin: Mechanistic insight into fitter strain propagation, Sci. Rep. 1 (2011) 531–569. doi:10.1038/srep00200.

[158] T.R. Maines, L.-M. Chen, N. Van Hoeven, T.M. Tumpey, O. Blixt, J.A. Belser, K.M. Gustin, M.B. Pearce, C. Pappas, J. Stevens, N.J. Cox, J.C. Paulson, R. Raman, R. Sasisekharan, J.M. Katz, R.O. Donis, Effect of receptor binding domain mutations on receptor binding and transmissibility of avian influenza H5N1 viruses., Virology. 413 (2011) 139–47. doi:10.1016/j.virol.2011.02.015.

[159] G. Neumann, C.A. Macken, A.I. Karasin, R.A.M. Fouchier, Y. Kawaoka, Egyptian H5N1 influenza viruses-cause for concern?, PLoS Pathog. 8 (2012) e1002932. doi:10.1371/journal.ppat.1002932.

[160] C.A. Russell, J.M. Fonville, A.E.X. Brown, D.F. Burke, D.L. Smith, S.L. James, S. Herfst, S. van Boheemen, M. Linster, E.J. Schrauwen, L. Katzelnick, A. Mosterín,

T. Kuiken, E. Maher, G. Neumann, A.D.M.E. Osterhaus, Y. Kawaoka, R.A.M. Fouchier, D.J. Smith, The Potential for Respiratory Droplet–Transmissible A/H5N1 Influenza Virus to Evolve in a Mammalian Host, Science (80-.). 336 (2012).

[161] K.G. Mansfield, Viral tropism and the pathogenesis of influenza in the Mammalian host., Am. J. Pathol. 171 (2007) 1089–92. doi:10.2353/ajpath.2007.070695.

[162] M.N. Matrosovich, T.Y. Matrosovich, T. Gray, N.A. Roberts, H.-D. Klenk, Human and avian influenza viruses target different cell types in cultures of human airway epithelium., Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 4620–4. doi:10.1073/pnas.0308001101.

[163] K. Shinya, M. Ebina, S. Yamada, M. Ono, N. Kasai, Y. Kawaoka, Avian flu: influenza virus receptors in the human airway., Nature. 440 (2006). doi:10.1038/440435a.

[164] R. Gao, B. Cao, Y. Hu, Z. Feng, D. Wang, W. Hu, J. Chen, Z. Jie, H. Qiu, K. Xu, X. Xu, H. Lu, W. Zhu, Z. Gao, N. Xiang, Y. Shen, Z. He, Y. Gu, Z. Zhang, Y. Yang, X. Zhao, L. Zhou, X. Li, S. Zou, Y. Zhang, X. Li, L. Yang, J. Guo, J. Dong, Q. Li, L. Dong, Y. Zhu, T. Bai, S. Wang, P. Hao, W. Yang, Y. Zhang, J. Han, H. Yu, D. Li, G.F. Gao, G. Wu, Y. Wang, Z. Yuan, Y. Shu, Human infection with a novel avian-origin influenza A (H7N9) virus., N. Engl. J. Med. 368 (2013) 1888–97. doi:10.1056/NEJMoa1304459.

[165] Q. Li, L. Zhou, M. Zhou, Z. Chen, F. Li, H. Wu, N. Xiang, E. Chen, F. Tang, D. Wang, L. Meng, Z. Hong, W. Tu, Y. Cao, L. Li, F. Ding, B. Liu, M. Wang, R. Xie, R. Gao, X. Li, T. Bai, S. Zou, J. He, J. Hu, Y. Xu, C. Chai, S. Wang, Y. Gao, L. Jin, Y. Zhang, H. Luo, H. Yu, L. Gao, X. Pang, G. Liu, Y. Shu, W. Yang, T.M. Uyeki, Y. Wang, F. Wu, Z. Feng, Preliminary Report: Epidemiology of the Avian Influenza A (H7N9) Outbreak in China., N. Engl. J. Med. (2013) 1–11. doi:10.1056/NEJMoa1304617.

[166] T.Y. Kwon, S.S. Lee, C.Y. Kim, J.Y. Shin, S.Y. Sunwoo, Y.S. Lyoo, Genetic characterization of H7N2 influenza virus isolated from pigs., Vet. Microbiol. 153 (2011) 393–7. doi:10.1016/j.vetmic.2011.06.011.

[167] M. Hirst, C.R. Astell, M. Griffith, S.M. Coughlin, M. Moksa, T. Zeng, D.E. Smailus, R.A. Holt, S. Jones, M.A. Marra, M. Petric, M. Krajden, D. Lawrence, A. Mak, R. Chow, D.M. Skowronski, S.A. Tweed, S. Goh, R.C. Brunham, J. Robinson, V. Bowes, K. Sojonky, S.K. Byrne, Y. Li, D. Kobasa, T. Booth, M. Paetzel, Novel avian influenza H7N3 strain outbreak, British Columbia., Emerg. Infect. Dis. 10 (2004) 2192–5. doi:10.3201/eid1012.040743.

[168] N. Van Hoeven, C. Pappas, J.A. Belser, T.R. Maines, H. Zeng, A. García-Sastre, R. Sasisekharan, J.M. Katz, T.M. Tumpey, Human HA and polymerase subunit PB2 proteins confer transmission of an avian influenza virus through the air., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 3366–71. doi:10.1073/pnas.0813172106.

[169] J.J. Skehel, D.C. Wiley, Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin., Annu. Rev. Biochem. 69 (2000) 531–69. doi:10.1146/annurev.biochem.69.1.531.

[170] J.M. Nicholls, A.J. Bourne, H. Chen, Y. Guan, J.M. Peiris, G. Lamblin, M. Lhermitte, A. Klein, P. Roussel, H. Van Halbeek, J. Vliegenthart, M. Matrosovich, T. Matrosovich, T. Gray, N. Roberts, H. Klenk, L. Baum, J. Paulson, A. Cerna, P. Janega, P. Martanovic, M. Lisy, P. Babal, P. Delmotte, S. Degroote, M. Merten, I. Van Seuningen, A. Bernigaud, C. Figarella, P. Roussel, J. Perini, A. Barkhordari, R. Stoddart, S. McClure, J. McClure, K. Shinya, M. Ebina, S. Yamada, M. Ono, N. Kasai, Y. Kawaoka, G. Rogers, J. Paulson, T. Ito, J. Couceiro, S. Kelm, L. Baum, S. Krauss, M. Castrucci, I. Donatelli, H. Kida, J. Paulson, R. Webster, Y. Kawaoka, Y. Suzuki, D. Mason, K. Micklem, M. Jones, Y. Konami, K. Yamamoto, T. Osawa, T. Irimura, R. Wagner, M. Matrosovich, H. Klenk, J. Skehel, D. Wiley, K. Shinya, M. Hatta, S. Yamada, A. Takada, S. Watanabe, P. Halfmann, T. Horimoto, G. Neumann, J. Kim, W. Lim, Y. Guan, M. Peiris, M. Kiso, T. Suzuki, Y. Suzuki, Y. Kawaoka, P. Gagneux, M. Cheriyan, N. Hurtado-Ziola, E. van der Linden, D. Anderson, H. McClure, A. Varki, N. Varki, S. Shi, R. Cote, C. Taylor, F. Dall'Olio, M. Chiricolo, A. D'Errico, E. Gruppioni, A. Altimari, M. Fiorentino, W. Grigioni, A. Gambaryan, S. Yamnikova, D. Lvov, A. Tuzikov, A. Chinarev, G.

Pazynina, R. Webster, M. Matrosovich, N. Bovin, J. Nicholls, M. Chan, W. Chan, H. Wong, C. Cheung, D. Kwong, M. Wong, W. Chui, L. Poon, S. Tsao, Y. Guan, J. Peiris, A. Ibricevic, A. Pekosz, M. Walter, C. Newby, J. Battaile, E. Brown, M. Holtzman, S. Brody, Sialic acid receptor detection in the human respiratory tract: evidence for widespread distribution of potential binding sites for human and avian influenza viruses, Respir. Res. 8 (2007) 73. doi:10.1186/1465-9921-8-73.

[171] D. Liu, W. Shi, Y. Shi, D. Wang, H. Xiao, W. Li, Y. Bi, Y. Wu, X. Li, J. Yan, W. Liu, G. Zhao, W. Yang, Y. Wang, J. Ma, Y. Shu, F. Lei, G.F. Gao, Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses., Lancet (London, England). 381 (2013) 1926–32. doi:10.1016/S0140-6736(13)60938-1.

[172] A. Jayaraman, A. Chandrasekaran, K. Viswanathan, R. Raman, J.G. Fox, R. Sasisekharan, Decoding the distribution of glycan receptors for human-adapted influenza A viruses in ferret respiratory tract., PLoS One. 7 (2012) e27517. doi:10.1371/journal.pone.0027517.

[173] A. Chandrasekaran, A. Srinivasan, R. Raman, K. Viswanathan, S. Raguram, T.M. Tumpey, V. Sasisekharan, R. Sasisekharan, Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin, Nat. Biotechnol. 26 (2008) 107–113. doi:10.1038/nbt1375.

[174] D. van Riel, V.J. Munster, E. de Wit, G.F. Rimmelzwaan, R.A.M. Fouchier, A.D.M.E. Osterhaus, T. Kuiken, Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals., Am. J. Pathol. 171 (2007) 1215–23. doi:10.2353/ajpath.2007.070248.

[175] W. Li, W. Shi, H. Qiao, S.Y.W. Ho, A. Luo, Y. Zhang, C. Zhu, Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses., Virol. J. 8 (2011) 183. doi:10.1186/1743-422X-8-183.

[176] K. Tharakaraman, R. Raman, N.W. Stebbins, K. Viswanathan, V. Sasisekharan, R. Sasisekharan, B.A. Cunha, C.J. Russell, R.G. Webster, J.K. Taubenberger, C. Scholtissek, W. Rohde, V. Von Hoyningen, R. Rott, Y. Kawaoka, S. Krauss, R.G. Webster, R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, Y. Kawaoka,

E.C. Settembre, P.R. Dormitzer, R. Rappuoli, J.K. Taubenberger, A.H. Reid, T.G. Fanning, J.K. Taubenberger, A.H. Reid, T.A. Janczewski, T.G. Fanning, O.T. Gorman, W.J. Bean, R.G. Webster, G.J. Nabel, C.J. Wei, J.E. Ledgerwood, G.J. Smith, J.L. Cherry, D.J. Lipman, A. Nikolskaya, Y.I. Wolf, D.J. Smith, C.A. Russell, A.G. Jansen, E.A. Sanders, A.W. Hoes, A.M. van Loon, E. Hak, M.K. Iwane, L. Simonsen, A.J. Caton, G.G. Brownlee, J.W. Yewdell, W. Gerhard, W. Gerhard, J. Yewdell, M.E. Frankel, R. Webster, J.C. Gaydos, F.H. Top, R.A. Hodder, P.K. Russell, D.E. Wentworth, M.W. McGregor, M.D. Macklin, V. Neumann, V.S. Hinshaw, E. Tsuchiya, D.C. Wiley, I.A. Wilson, J.J. Skehel, I.A. Wilson, J.J. Skehel, D.C. Wiley, Y. Bao, R. Xu, M. Zhang, S.J. Anthony, M.B. Pearce, D.J. Smith, S. Forrest, D.H. Ackley, A.S. Perelson, M.S. Lee, J.S. Chen, A. Klimov, L. Simonsen, K. Fukuda, N. Cox, G.W. Both, M.J. Sleigh, N.J. Cox, A.P. Kendal, M.T. Coiras, R.J. Connor, Y. Kawaoka, R.G. Webster, J.C. Paulson, A. Chandrasekaran, A. Srinivasan, J.C. de Jong, W.W. Thompson, M. Igarashi, K. Ito, H. Kida, A. Takada, Y.P. Lin, E.C. Settembre, P.R. Dormitzer, R. Rappuoli, Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic, Sci. Rep. 3 (2013) 141–155. doi:10.1038/srep01822.

[177] K. Srinivasan, R. Raman, A. Jayaraman, K. Viswanathan, R. Sasisekharan, Quantitative description of glycan-receptor binding of influenza A virus H7 hemagglutinin., PLoS One. 8 (2013) e49597. doi:10.1371/journal.pone.0049597.

[178] C.J. Russell, R.G. Webster, The genesis of a pandemic influenza virus, Cell. 123 (2005) 368–371. doi:10.1016/j.cell.2005.10.019.

[179] J. Taubenberger, A. Reid, R. Lourens, R. Wang, G. Jin, T. Fanning, Characterization of the 1918 influenza virus polymerase genes., Nature. 437 (2005) 889–93. doi:10.1038/nature04230.

[180] C. Scholtissek, W. Rohde, V. Von Hoyningen, R. Rott, On the origin of the human influenza virus subtypes H2N2 and H3N2, Virology. 87 (1978) 13–20. doi:10.1016/0042-6822(78)90153-8.

[181] Y. Kawaoka, S. Krauss, R.G. Webster, Avian-to-human transmission of the PB1

gene of influenza A viruses in the 1957 and 1968 pandemics., J. Virol. 63 (1989) 4603–8. http://www.ncbi.nlm.nih.gov/pubmed/2795713 (accessed November 14, 2016).

[182] et al. Webster, R. G., Evolution and ecology of influenza A viruses, Microbiol Rev. 56 (1992) 152–179.

[183] E.C. Settembre, P.R. Dormitzer, R. Rappuoli, H1N1: can a pandemic cycle be broken?, Sci Transl Med. 2 (2010) 24ps14. doi:10.1126/scitranslmed.3000948.

[184] J.K. Taubenberger, A.H. Reid, T.G. Fanning, The 1918 influenza virus: A killer comes into view., Virology. 274 (2000) 241–245. doi:10.1006/viro.2000.0495.

[185] J.K. Taubenberger, A.H. Reid, T.A. Janczewski, T.G. Fanning, Integrating historical, clinical and molecular genetic data in order to explain the origin and virulence of the 1918 Spanish influenza virus., Philos. Trans. R. Soc. Lond. B. Biol. Sci. 356 (2001) 1829–1839. doi:10.1098/rstb.2001.1020.

[186] O.T. Gorman, W.J. Bean, R.G. Webster, Evolutionary processes in influenza viruses: divergence, rapid evolution, and stasis., Curr. Top. Microbiol. Immunol. 176 (1992) 75–97. doi:10.1007/978-3-642-77011-1.

[187] G.J. Nabel, C.-J. Wei, J.E. Ledgerwood, Vaccinate for the next H2N2 pandemic now., Nature. 471 (2011) 157–158. doi:10.1038/471157a.

[188] J.L. Cherry, D.J. Lipman, A. Nikolskaya, Y. Wolf, Evolutionary dynamics of N-Glycosylation sites of influenza virus Hemagglutinin, PLoS Curr. 1 (2009) RRN1001. doi:10.1371/currents.RRN1001.

[189] C.A. Russell, T.C. Jones, I.G. Barr, N.J. Cox, R.J. Garten, V. Gregory, I.D. Gust, A.W. Hampson, A.J. Hay, A.C. Hurt, J.C. de Jong, A. Kelso, A.I. Klimov, T. Kageyama, N. Komadina, A.S. Lapedes, Y.P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A.D.M.E. Osterhaus, G.F. Rimmelzwaan, M.W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R.A.M. Fouchier, D.J. Smith, The global circulation of seasonal influenza A (H3N2) viruses., Science. 320 (2008) 340–346. doi:10.1126/science.1154137.

[190] A.G.S.C. Jansen, E.A.M. Sanders, A.W. Hoes, A.M. Van Loon, E. Hak, Influenza-and respiratory syncytial virus-associated mortality and hospitalisations, Eur.

Respir. J. 30 (2007) 1158–1166. doi:10.1183/09031936.00034407.

[191] M.K. Iwane, K.M. Edwards, P.G. Szilagyi, F.J. Walker, M.R. Griffin, G.A. Weinberg, C. Coulen, K.A. Poehling, L.P. Shone, S. Balter, C.B. Hall, D.D. Erdman, K. Wooten, B. Schwartz, Population-based surveillance for hospitalizations associated with respiratory syncytial virus, influenza virus, and parainfluenza viruses among young children, Pediatrics. 113 (2004) 1758–1764. doi:10.1542/peds.113.6.1758.

[192] L. Simonsen, T.A. Reichert, C. Viboud, W.C. Blackwelder, R.J. Taylor, M.A. Miller, Impact of influenza vaccination on seasonal mortality in the US elderly population, Arch Intern Med. 165 (2005) 265–272. doi:10.1001/archinte.165.3.265.

[193] A. Srinivasan, K. Viswanathan, R. Raman, A. Chandrasekaran, S. Raguram, T.M. Tumpey, V. Sasisekharan, R. Sasisekharan, Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 2800–5. doi:10.1073/pnas.0711963105.

[194] A.J. Caton, G.G. Brownlee, J.W. Yewdell, W. Gerhard, The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype), Cell. 31 (1982) 417–427. doi:10.1016/0092-8674(82)90135-0.

[195] W. Gerhard, J. Yewdell, M.E. Frankel, R. Webster, Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies., Nature. 290 (1981) 713–717. doi:10.1038/290713a0.

[196] J.C. Gaydos, F.H. Top, R.A. Hodder, P.K. Russell, Swine influenza A outbreak, Fort Dix, New Jersey, 1976, Emerg. Infect. Dis. 12 (2006) 23–28. doi:10.3201/eid1201.050965.

[197] D.E. Wentworth, M.W. McGregor, M.D. Macklin, V. Neumann, V.S. Hinshaw, Transmission of swine influenza virus to humans after exposure to experimentally infected pigs, J Infect Dis. 175 (1997) 7–15. http://www.ncbi.nlm.nih.gov/pubmed/8985190 (accessed November 14, 2016).

[198] E. Tsuchiya, K. Sugawara, S. Hongo, Y. Matsuzaki, Y. Muraki, Z.N. Li, K. Nakamura, Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus, J. Gen. Virol. 82 (2001) 2475–2484. doi:10.1099/0022-1317-82-10-2475.

[199] D.C. Wiley, I.A. Wilson, J.J. Skehel, Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation., Nature. 289 (1981) 373–378. doi:10.1038/289373a0.

[200] M. Zhang, B. Gaschen, W. Blay, B. Foley, N. Haigwood, C. Kuiken, B. Korber, Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin, Glycobiology. 14 (2004) 1229–1246. doi:10.1093/glycob/cwh106.

[201] I. Navarrete-Macias, R. Rabadan, J. Pedersen, J.M. Chan, W. Karesh, W.I. Lipkin, K. Pugliares, M. Sanchez-Leon, P. Daszak, Z.W. Carpenter, J.T. Saliki, H.S. Ip, J.A. St. Leger, S.J. Anthony, T. Rowles, Emergence of Fatal Avian Influenza in New England Harbor Seals, MBio. 3 (2012) e00166-12-e00166-12. doi:10.1128/mBio.00166-12.

[202] M.B. Pearce, A. Jayaraman, C. Pappas, J.A. Belser, H. Zeng, K.M. Gustin, T.R. Maines, X. Sun, R. Raman, N.J. Cox, R. Sasisekharan, J.M. Katz, T.M. Tumpey, Pathogenesis and transmission of swine origin A(H3N2)v influenza viruses in ferrets., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 3944–9. doi:10.1073/pnas.1119945109.

[203] D.J. Smith, S. Forrest, D.H. Ackley, A.S. Perelson, Variable efficacy of repeated annual influenza vaccination., Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 14001–6. doi:10.1073/pnas.96.24.14001.

[204] M.S. Lee, J.S.E. Chen, Predicting antigenic variants of influenza A/H3N2 viruses, Emerg. Infect. Dis. 10 (2004) 1385–1390. doi:10.3201/eid1008.040107.

[205] A. Klimov, L. Simonsen, K. Fukuda, N. Cox, Surveillance and impact of influenza in the United States, J. Am. Geriatr. Soc. 17 (1999) 42–46. http://www.academia.edu/20793601/Surveillance_and_impact_of_influenza_in_the_United_States.

[206] WHO Recommended composition of influenza virus vaccines for use in the 2013 southern hemisphere influenza season., Wkly Epidemiol Rec. (2012) 389–400.

[207] G.W. Both, M.J. Sleigh, N.J. Cox, A.P. Kendal, Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential

amino acid changes at key antigenic sites., J. Virol. 48 (1983) 52–60. http://www.ncbi.nlm.nih.gov/pubmed/6193288 (accessed November 14, 2016).

[208] M.T. Coiras, J.C. Aguilar, M. Galiano, S. Carlos, V. Gregory, Y.P. Lin, A. Hay, P. Pérez-Breña, Rapid molecular analysis of the haemagglutinin gene of human influenza A H3N2 viruses isolated in Spain from 1996 to 2000, Arch. Virol. 146 (2001) 2133–2147. doi:10.1007/s007050170025.

[209] R.J. Connor, Y. Kawaoka, R.G. Webster, J.C. Paulson, Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates., Virology. 205 (1994) 17–23. doi:10.1006/viro.1994.1615.

[210] J.C. de Jong, D.J. Smith, A.S. Lapedes, I. Donatelli, L. Campitelli, G. Barigazzi, K. Van Reeth, T.C. Jones, G.F. Rimmelzwaan, A.D.M.E. Osterhaus, R.A.M. Fouchier, Antigenic and genetic evolution of swine influenza A (H3N2) viruses in Europe., J. Virol. 81 (2007) 4315–4322. doi:10.1128/JVI.02458-06.

[211] W.W. Thompson, D.K. Shay, E. Weintraub, L. Brammer, N. Cox, L.J. Anderson, K. Fukuda, Mortality associated with influenza and respiratory syncytial virus in the United States, JAMA. 289 (2003) 179–186. doi:joc21709 [pii].

[212] M. Igarashi, K. Ito, H. Kida, A. Takada, Genetically destined potentials for N-linked glycosylation of influenza virus hemagglutinin, Virology. 376 (2008) 323–329. doi:10.1016/j.virol.2008.03.036.

[213] S.J. Anthony, J.A. St Leger, K. Pugliares, H.S. Ip, J.M. Chan, Z.W. Carpenter, I. Navarrete-Macias, M. Sanchez-Leon, J.T. Saliki, J. Pedersen, W. Karesh, P. Daszak, R. Rabadan, T. Rowles, W.I. Lipkin, Emergence of fatal avian influenza in New England harbor seals., MBio. 3 (2012) e00166-12. doi:10.1128/mBio.00166-12.

[214] E. a Karlsson, H.S. Ip, J.S. Hall, S.W. Yoon, J. Johnson, M. a Beck, R.J. Webby, S. Schultz-Cherry, Respiratory transmission of an avian H3N8 influenza virus isolated from a harbour seal., Nat. Commun. 5 (2014) 4791. doi:10.1038/ncomms5791.

[215] B. Adamczyk, T. Tharmalingam, P.M. Rudd, Glycans as cancer biomarkers., Biochim. Biophys. Acta. 1820 (2012) 1347–1353.

doi:10.1016/j.bbagen.2011.12.001.

[216] X.-E. Liu, L. Desmyter, C.-F. Gao, W. Laroy, S. Dewaele, V. Vanhooren, L. Wang, H. Zhuang, N. Callewaert, C. Libert, R. Contreras, C. Chen, N-glycomic changes in hepatocellular carcinoma patients with liver cirrhosis induced by hepatitis B virus., Hepatology. 46 (2007) 1426–1435. doi:10.1002/hep.21855.

[217] M. Guerrini, Z. Zhang, Z. Shriver, A. Naggi, S. Masuko, R. Langer, B. Casu, R.J. Linhardt, G. Torri, R. Sasisekharan, Orthogonal analytical approaches to detect potential contaminants in heparin., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 16956–16961. doi:10.1073/pnas.0906861106.

[218] C.A. Waddling, T.H. Plummer, A.L. Tarentino, P. Van Roey, Structural basis for the substrate specificity of endo-beta-N-acetylglucosaminidase F(3)., Biochemistry. 39 (2000) 7878–7885. http://www.ncbi.nlm.nih.gov/pubmed/10891067.

[219] T. Patel, J. Bruce, A. Merry, C. Bigge, M. Wormald, A. Jaques, R. Parekh, Use of hydrazine to release in intact and unreduced form both N- and O-linked oligosaccharides from glycoproteins., Biochemistry. 32 (1993) 679–693. http://www.ncbi.nlm.nih.gov/pubmed/8422375.

[220] D. Aminoff, W. Gathmann, Quantitation of Oligosaccharides Released by B-Elimination reaction, Anal. Biochem. 53 (1980) 44–53. http://www.sciencedirect.com/science/article/pii/000326978090038X (accessed May 16, 2013).

[221] D.J. Harvey, Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates., Mass Spectrom. Rev. 18 (1999) 349–450. doi:10.1002/(SICI)1098-2787(1999)18:6<349::AID-MAS1>3.0.CO;2-H.

[222] Y. Wada, P. Azadi, C.E. Costello, A. Dell, R.A. Dwek, H. Geyer, R. Geyer, K. Kakehi, N.G. Karlsson, K. Kato, N. Kawasaki, K.-H. Khoo, S. Kim, A. Kondo, E. Lattova, Y. Mechref, E. Miyoshi, K. Nakamura, H. Narimatsu, M. V Novotny, N.H. Packer, H. Perreault, J. Peter-Katalinic, G. Pohlentz, V.N. Reinhold, P.M. Rudd, A. Suzuki, N. Taniguchi, Comparison of the methods for profiling glycoprotein glycans--HUPO Human Disease Glycomics/Proteome Initiative multi-institutional

study., Glycobiology. 17 (2007) 411–422. doi:10.1093/glycob/cwl086.

[223] Y. Mechref, Y. Hu, A. Garcia, A. Hussein, Identifying cancer biomarkers by mass spectrometry-based glycomics., Electrophoresis. 33 (2012) 1755–1767. doi:10.1002/elps.201100715.

[224] A. Ceroni, K. Maass, H. Geyer, GlycoWorkbench : A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans, J. Proteome Res. 7 (2008) 1650–1659. http://pubs.acs.org/doi/abs/10.1021/pr7008252 (accessed May 15, 2013).

[225] R. Raman, M. Venkataraman, S. Ramakrishnan, W. Lang, S. Raguram, R. Sasisekharan, Advancing glycomics: implementation strategies at the consortium for functional glycomics., Glycobiology. 16 (2006) 82R–90R. doi:10.1093/glycob/cwj080.

[226] H.-J. Gabius, S. André, J. Jiménez-Barbero, A. Romero, D. Solís, From lectin structure to functional glycomics: principles of the sugar code., Trends Biochem. Sci. 36 (2011) 298–313. doi:10.1016/j.tibs.2011.01.005.

[227] J. Hirabayashi, M. Yamada, A. Kuno, H. Tateno, Lectin microarrays: concept, principle and applications., Chem. Soc. Rev. 42 (2013) 4443–4458. doi:10.1039/c3cs35419a.

[228] J. Hirabayashi, Lectin-based structural glycomics: glycoproteomics and glycan profiling., Glycoconj. J. 21 (2004) 35–40. doi:10.1023/B:GLYC.0000043745.18988.a1.

[229] K.T. Pilobello, D.E. Slawek, L.K. Mahal, A ratiometric lectin microarray approach to analysis of the dynamic mammalian glycome., Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 11534–11539. doi:10.1073/pnas.0704954104.

[230] A. V Nairn, W.S. York, K. Harris, E.M. Hall, J.M. Pierce, K.W. Moremen, Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes., J. Biol. Chem. 283 (2008) 17298–313. doi:10.1074/jbc.M801964200.

[231] M.M. Winslow, T.L. Dayton, R.G.W. Verhaak, C. Kim-Kiselak, E.L. Snyder, D.M. Feldser, D.D. Hubbard, M.J. DuPage, C. a Whittaker, S. Hoersch, S. Yoon, D.

Crowley, R.T. Bronson, D.Y. Chiang, M. Meyerson, T. Jacks, Suppression of lung adenocarcinoma progression by Nkx2-1., Nature. 473 (2011) 101–4. doi:10.1038/nature09881.

[232] N.E. Reticker-Flynn, D.F.B. Malta, M.M. Winslow, J.M. Lamar, M.J. Xu, G.H. Underhill, R.O. Hynes, T.E. Jacks, S.N. Bhatia, A combinatorial extracellular matrix platform identifies cell-extracellular matrix interactions that correlate with metastasis., Nat. Commun. 3 (2012) 1122. doi:10.1038/ncomms2128.

[233] S.S. Pinho, C.A. Reis, Glycosylation in cancer: mechanisms and clinical implications, Nat. Rev. Cancer. 15 (2015) 540–555. doi:10.1038/nrc3982.

[234] I.O. Potapenko, V.D. Haakensen, T. Lüders, A. Helland, I. Bukholm, T. Sørlie, V.N. Kristensen, O.C. Lingjaerde, A.-L. Børresen-Dale, Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression., Mol. Oncol. 4 (2010) 98–118. doi:10.1016/j.molonc.2009.12.001.

[235] M. Takahashi, Y. Kuroki, K. Ohtsubo, N. Taniguchi, Core fucose and bisecting GlcNAc, the direct modifiers of the N-glycan core: their functions and target proteins., Carbohydr. Res. 344 (2009) 1387–90. doi:10.1016/j.carres.2009.04.031.

[236] Y. Zhao, S. Itoh, X. Wang, T. Isaji, E. Miyoshi, Y. Kariya, K. Miyazaki, N. Kawasaki, N. Taniguchi, J. Gu, Deletion of Core Fucosylation on alpha3beta1 Integrin Down-regulates Its Functions, J. Biol. Chem. 281 (2006) 38343–38350. doi:10.1074/jbc.M608764200.

[237] Y.-C. Liu, H.-Y. Yen, C.-Y. Chen, C.-H. Chen, P.-F. Cheng, Y.-H. Juan, C.-H. Chen, K.-H. Khoo, C.-J. Yu, P.-C. Yang, T.-L. Hsu, C.-H. Wong, Sialylation and fucosylation of epidermal growth factor receptor suppress its dimerization and activation in lung cancer cells., Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 11332–11337. doi:10.1073/pnas.1107385108.

[238] P. Radhakrishnan, S. Dabelsteen, F.B. Madsen, C. Francavilla, K.L. Kopp, C. Steentoft, S.Y. Vakhrushev, J. V. Olsen, L. Hansen, E.P. Bennett, A. Woetmann, G. Yin, L. Chen, H. Song, M. Bak, R.A. Hlady, S.L. Peters, R. Opavsky, C. Thode, K. Qvortrup, K.T.-B.G. Schjoldager, H. Clausen, M.A. Hollingsworth, H.H.

Wandall, Immature truncated O-glycophenotype of cancer directly induces oncogenic features, Proc. Natl. Acad. Sci. 111 (2014) E4066–E4075. doi:10.1073/pnas.1406619111.

[239] S.S. Pinho, C.A. Reis, Glycosylation in cancer: mechanisms and clinical implications, Nat Rev Cancer. 15 (2015) 540–555. doi:10.1038/nrc3982.

[240] R. Siegel, K. Miller, A. Jemal, Cancer statistics , 2015 ., CA Cancer J Clin. 65 (2015) 29. doi:10.3322/caac.21254.

[241] C.K. Anders, L. a Carey, Biology, Metastatic Patterns and Treatment of Patients with Triple-Negtive Breast Cancer, Breast. 9 (2010) S73–S81. doi:10.3816/CBC.2009.s.008.Biology.

[242] M.A. Frese, F. Milz, M. Dick, W.C. Lamanna, T. Dierks, Characterization of the human sulfatase Sulf1 and its high affinity heparin/heparan sulfate interaction domain, J. Biol. Chem. 284 (2009) 28033–28044. doi:10.1074/jbc.M109.035808.

[243] Y. Hu, G.K. Smyth, ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays, J. Immunol. Methods. 347 (2009) 70–78. doi:10.1016/j.jim.2009.06.008.

[244] B. Elenbaas, L. Spirio, F. Koerner, M.D. Fleming, D.B. Zimonjic, J.L. Donaher, N.C. Popescu, W.C. Hahn, R. a Weinberg, Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells., Genes Dev. 15 (2001) 50–65. doi:10.1101/gad.828901.monly.

[245] W.C. Hahn, S.K. Dessain, M.W. Brooks, J.E. King, B. Elenbaas, D.M. Sabatini, J.A. DeCaprio, R.A. Weinberg, Enumeration of the simian virus 40 early region elements necessary for human cell transformation, Mol Cell Biol. 22 (2002) 2111–2123. doi:10.1128/MCB.22.7.2111.

[246] G. Dontu, W.M. Abdallah, J.M. Foley, K.W. Jackson, M.F. Clarke, M.J. Kawamura, M.S. Wicha, In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells., Genes Dev. 17 (2003) 1253–70. doi:10.1101/gad.1061803.

[247] N.C. Shaner, R.E. Campbell, P.A. Steinbach, B.N.G. Giepmans, A.E. Palmer, R.Y. Tsien, Improved monomeric red, orange and yellow fluorescent proteins

derived from Discosoma sp. red fluorescent protein., Nat. Biotechnol. 22 (2004) 1567–72. doi:10.1038/nbt1037.

[248] C.L. Chaffer, R.A. Weinberg, How does multistep tumorigenesis really proceed?, Cancer Discov. 5 (2015) 22–24. doi:10.1158/2159-8290.CD-14-0788.

[249] S.Y. Park, H.E. Lee, H. Li, M. Shipitsin, R. Gelman, K. Polyak, Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer, Clin. Cancer Res. 16 (2010) 876–887. doi:10.1158/1078-0432.CCR-09-1532.

[250] C.L. Chaffer, N.D. Marjanovic, T. Lee, G. Bell, C.G. Kleer, F. Reinhardt, A.C. D'Alessio, R.A. Young, R.A. Weinberg, Poised chromatin at the ZEB1 promoter in breast cancer enables cell plasticity and enhances tumorigenicity, Cell. in press (2013).

[251] J.C. Kathryn, G. Sireesha V, L. Stanley, Triple Negative Breast Cancer Cell Lines: One Tool in the Search for Better Treatment of Triple Negative Breast Cancer, Breast Dis. 32 (2012) 35–48. doi:10.3233/BD-2010-0307.Triple.

[252] T.A. Ince, A.L. Richardson, G.W. Bell, M. Saitoh, S. Godar, A.E. Karnoub, J.D. Iglehart, R.A. Weinberg, Transformation of Different Human Breast Epithelial Cell Types Leads to Distinct Tumor Phenotypes, Cancer Cell. 12 (2007) 160–170. doi:10.1016/j.ccr.2007.06.013.

[253] C. Sheridan, H. Kishimoto, R.K. Fuchs, S. Mehrotra, P. Bhat-Nakshatri, C.H. Turner, R. Goulet, S. Badve, H. Nakshatri, CD44+/CD24- breast cancer cells exhibit enhanced invasive properties: an early step necessary for metastasis., Breast Cancer Res. 8 (2006) 59. doi:10.1186/bcr1610.

[254] M. Al-Hajj, M.F. Clarke, Self-renewal and solid tumor stem cells., Oncogene. 23 (2004) 7274–82. doi:10.1038/sj.onc.1207947.

[255] J. Kanodia, D. Chai, J. Vollmer, J. Kim, A. Raue, G. Finn, B. Schoeberl, Deciphering the mechanism behind Fibroblast Growth Factor (FGF) induced biphasic signal-response profiles., Cell Commun. Signal. 12 (2014) 34. doi:10.1186/1478-811X-12-34.

[256] K.E. Sung, X. Su, E. Berthier, C. Pehlke, A. Friedl, D.J. Beebe, Understanding the

Impact of 2D and 3D Fibroblast Cultures on In Vitro Breast Cancer Models, PLoS One. 8 (2013) 1–13. doi:10.1371/journal.pone.0076373.

[257] R. Singhai, V.W. Patil, S.R. Jaiswal, S.D. Patil, M.B. Tayade, A. V Patil, E-Cadherin as a diagnostic biomarker in breast cancer., N. Am. J. Med. Sci. 3 (2011) 227–33. doi:10.4297/najms.2011.3227.

[258] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, A.M. Chinnaiyan, ONCOMINE: a cancer microarray database and integrated data-mining platform, ... (New York, NY). 6 (2004) 1–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1635162/ (accessed May 30, 2013).

[259] A.M. Szász, A. Lánczky, Á. Nagy, S. Förster, K. Hark, J.E. Green, A. Boussioutas, R. Busuttil, A. Szabó, B. Győrffy, A.M. Szász, A. Lánczky, Á. Nagy, S. Förster, K. Hark, J.E. Green, A. Boussioutas, R. Busuttil, A. Szabó, B. Győrffy, Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients, Oncotarget. 7 (2016) 49322–49333. doi:10.18632/oncotarget.10337.

[260] R. Dhanasekaran, I. Nakamura, C. Hu, G. Chen, A.M. Oseini, E.S. Seven, A.G. Miamen, C.D. Moser, W. Zhou, T.H. van Kuppevelt, J.M. van Deursen, T. Mounajjed, M.E. Fernandez-Zapico, L.R. Roberts, Activation of the transforming growth factor-β/SMAD transcriptional pathway underlies a novel tumor-promoting role of sulfatase 1 in hepatocellular carcinoma, Hepatology. 61 (2015) 1269–1283. doi:10.1002/hep.27658.

[261] E. Von Mutius, D. Vercelli, M.E. von, E. Von Mutius, D. Vercelli, Farm living: effects on childhood asthma and allergy, Nat. Rev. Immunol. 10 (2010) 861–868. doi:10.1038/nri2871.

[262] M. Masoli, D. Fabian, S. Holt, R. Beasley, The global burden of asthma: Executive summary of the GINA Dissemination Committee Report, Allergy Eur. J. Allergy Clin. Immunol. 59 (2004) 469–478. doi:10.1111/j.1398-9995.2004.00526.x.

[263] M.J. Ege, M. Mayer, A.-C. Normand, J. Genuneit, W.O.C.M. Cookson, C. Braun-Fahrländer, D. Heederik, R. Piarroux, E. von Mutius, Exposure to Environmental Microorganisms and Childhood Asthma, N. Engl. J. Med. 364 (2011) 701–709.

doi:10.1056/NEJMoa1007302.

[264] M.J. Schuijs, M.A. Willart, K. Vergote, D. Gras, K. Deswarte, M.J. Ege, F.B. Madeira, R. Beyaert, G. van Loo, F. Bracher, E. von Mutius, P. Chanez, B.N. Lambrecht, H. Hammad, Farm dust and endotoxin protect against allergy through A20 induction in lung epithelial cells., Science. 349 (2015) 1106–10. doi:10.1126/science.aac6623.

[265] M.M. Stein, C.L. Hrusch, J. Gozdz, C. Igartua, V. Pivniouk, S.E. Murray, J.G. Ledford, M. Marques dos Santos, R.L. Anderson, N. Metwali, J.W. Neilson, R.M. Maier, J.A. Gilbert, M. Holbreich, P.S. Thorne, F.D. Martinez, E. von Mutius, D. Vercelli, C. Ober, A.I. Sperling, Innate Immunity and Asthma Risk in Amish and Hutterite Farm Children, N. Engl. J. Med. 375 (2016) 411–421. doi:10.1056/NEJMoa1508749.

[266] M. Peters, M. Kauth, J. Schwarze, C. Körner-Rettberg, J. Riedler, D. Nowak, C. Braun-Fahrländer, E. von Mutius, a Bufe, O. Holst, Inhalation of stable dust extract prevents allergen induced airway inflammation and hyperresponsiveness., Thorax. 61 (2006) 134–139. doi:10.1136/thx.2005.049403.

[267] D.P. Strachan, Hay fever, hygiene, and household size, BMJ Br. Med. J. 299 (1989) 1259–1260. doi:10.1136/bmj.299.6710.1259.

[268] T. Alfvén, C. Braun-Fahrländer, B. Brunekreef, E. Von Mutius, J. Riedler, A. Scheynius, M. Van Hage, M. Wickman, M.R. Benz, J. Budde, K.B. Michels, D. Schram, E. Üblagger, M. Waser, G. Pershagen, Allergic diseases and atopic sensitization in children related to farming and anthroposophic lifestyle - The PARSIFAL study, Allergy Eur. J. Allergy Clin. Immunol. 61 (2006) 414–421. doi:10.1111/j.1398-9995.2005.00939.x.

[269] J. Riedler, C. Braun-Fahrländer, W. Eder, M. Schreuer, M. Waser, S. Maisch, D. Carr, R. Schierl, D. Nowak, E. von Mutius, A.S. Team*, Exposure to farming in early life and development of asthma and\rallergy: a cross-sectional survey, Lancet. 358 (2001) 1129–1133.

[270] C. Braun-Fahrländer, M. Gassner, L. Grize, U. Neu, F.H. Sennhauser, H.S. Varonier, J.C. Vuille, B. Wüthrich, Prevalence of hay fever and allergic

sensitization in farmer's children and their peers living in the same rural community, Clin. Exp. Allergy. 29 (1999) 28–34. doi:10.1046/j.1365-2222.1999.00479.x.

[271] S. Illi, M. Depner, J. Genuneit, E. Horak, G. Loss, C. Strunz-Lehner, G. Büchele, A. Boznanski, H. Danielewicz, P. Cullinan, D. Heederik, C. Braun-Fahrländer, E. Von Mutius, Protection from childhood asthma and allergy in Alpine farm environments - The GABRIEL Advanced Studies, J. Allergy Clin. Immunol. 129 (2012) 1470–1477. doi:10.1016/j.jaci.2012.03.013.

[272] M.J. Ege, R. Frei, C. Bieli, D. Schram-Bijkerk, M. Waser, M.R. Benz, G. Weiss, F. Nyberg, M. van Hage, G. Pershagen, B. Brunekreef, J. Riedler, R. Lauener, C. Braun-Fahrländer, E. von Mutius, Not all farming environments protect against the development of asthma and wheeze in children, J. Allergy Clin. Immunol. 119 (2007) 1140–1147. doi:10.1016/j.jaci.2007.01.037.

[273] M.R. Perkin, D.P. Strachan, Which aspects of the farming lifestyle explain the inverse association with childhood allergy?, J. Allergy Clin. Immunol. 117 (2006) 1374–1381. doi:10.1016/j.jaci.2006.03.008.

[274] J. Genuneit, Exposure to farming environments in childhood and asthma and wheeze in rural populations: A systematic review with meta-analysis, Pediatr. Allergy Immunol. 23 (2012) 509–518. doi:10.1111/j.1399-3038.2012.01312.x.

[275] B. Schaub, J. Liu, S. Höppler, I. Schleich, J. Huehn, S. Olek, G. Wieczorek, S. Illi, E. von Mutius, Maternal farm exposure modulates neonatal immune mechanisms through regulatory T cells, J. Allergy Clin. Immunol. 123 (2009). doi:10.1016/j.jaci.2009.01.056.

[276] P.I. Pfefferle, G. Büchele, N. Blümer, M. Roponen, M.J. Ege, S. Krauss-Etschmann, J. Genuneit, A. Hyvärinen, M.R. Hirvonen, R. Lauener, J. Pekkanen, J. Riedler, J.C. Dalphin, B. Brunekeef, C. Braun-Fahrländer, E. von Mutius, H. Renz, Cord blood cytokines are modulated by maternal farming activities and consumption of farm dairy products during pregnancy: The PASTURE Study, J. Allergy Clin. Immunol. 125 (2010). doi:10.1016/j.jaci.2009.09.019.

[277] J. Douwes, S. Cheng, N. Travier, C. Cohet, A. Niesink, J. McKenzie, C.

Cunningham, G. Le Gros, E. Von Mutius, N. Pearce, Farm exposure in utero may protect against asthma, hay fever and eczema, Eur. Respir. J. 32 (2008) 603–611. doi:10.1183/09031936.00033707.

[278] M.J. Ege, C. Bieli, R. Frei, R.T. van Strien, J. Riedler, E. Üblagger, D. Schram-Bijkerk, B. Brunekreef, M. van Hage, A. Scheynius, G. Pershagen, M.R. Benz, R. Lauener, E. von Mutius, C. Braun-Fahrländer, the PARSIFAL Study team, Prenatal farm exposure is related to the expression of receptors of the innate immunity and to atopic sensitization in school-age children, J. Allergy Clin. Immunol. 117 (2006) 817–823. doi:10.1016/j.jaci.2005.12.1307.

[279] J. Seedorf, J. Hartung, M. Schröder, K.H. Linkert, V.R. Phillips, M.R. Holden, R.W. Sneath, J.L. Short, R.P. White, S. Pedersen, H. Takai, J.O. Johnsen, J.H.M. Metz, P.W.G. Groot Koerkamp, G.H. Uenk, C.M. Wathes, Concentrations and Emissions of Airborne Endotoxins and Microorganisms in Livestock Buildings in Northern Europe, J. Agric. Eng. Res. 70 (1998) 97–109. doi:http://dx.doi.org/10.1006/jaer.1997.0281.

[280] K. Vogel, N. Blümer, M. Korthals, J. Mittelstädt, H. Garn, M. Ege, E. von Mutius, S. Gatermann, A. Bufe, T. Goldmann, K. Schwaiger, H. Renz, S. Brandau, J. Bauer, H. Heine, O. Holst, Animal shed Bacillus licheniformis spores possess allergy-protective as well as inflammatory properties, J. Allergy Clin. Immunol. 122 (2008). doi:10.1016/j.jaci.2008.05.016.

[281] C. Braun-Fahrländer, J. Riedler, U. Herz, W. Eder, M. Waser, L. Grize, S. Maisch, D. Carr, F. Gerlach, A. Bufe, R.P. Lauener, R. Schierl, H. Renz, D. Nowak, E. von Mutius, Environmental Exposure to Endotoxin and Its Relation to Asthma in School-Age Children, N. Engl. J. Med. 347 (2002) 869–877. doi:10.1056/NEJMoa020057.

[282] R.T. Van Strien, R. Engel, O. Holst, A. Bufe, W. Eder, M. Waser, C. Braun-Fahrländer, J. Riedler, D. Nowak, E. Von Mutius, Microbial exposure of rural school children, as assessed by levels of N-acetyl-muramic acid in mattress dust, and its association with respiratory health, J. Allergy Clin. Immunol. 113 (2004) 860–867. doi:10.1016/j.jaci.2004.01.783.

[283] H.D. Brightbill, R.L. Modlin, Toll-like receptors: Molecular mechanisms of the mammalian immune response, Immunology. 101 (2000) 1–10. doi:10.1046/j.1365-2567.2000.00093.x.

[284] K. Takeda, S. Akira, Toll-like receptors in innate immunity, Int. Immunol. 17 (2005) 1–14. doi:10.1093/intimm/dxh186.

[285] R.P. Lauener, T. Birchler, J. Adamski, Expression of CD14 and Toll-like receptor 2 in farmers ' and non- farmers ' children, Lancet. 360 (2002) 465–466.

[286] J. Debarry, H. Garn, A. Hanuszkiewicz, N. Dickgreber, N. Blümer, E. von Mutius, A. Bufe, S. Gatermann, H. Renz, O. Holst, H. Heine, Acinetobacter lwoffii and Lactococcus lactis strains isolated from farm cowsheds possess strong allergy-protective properties, J. Allergy Clin. Immunol. 119 (2007) 1514–1521. doi:10.1016/j.jaci.2007.03.023.

[287] USFDA, Guidance for Industry Botanical Drug Products, U.S. Dep. Heal. Hum. Serv. Food Drug Adm. Cent. Drug Eval. Res. (2004) 1–47.

[288] J. Anderson, C. Bell, J. Bishop, I. Capila, T. Ganguly, J. Glajch, M. Iyer, G. Kaundinya, J. Lansing, J. Pradines, J. Prescott, B.A. Cohen, D. Kantor, R. Sachleben, Demonstration of equivalence of a generic glatiramer acetate (Glatopa), J. Neurol. Sci. 359 (2015) 24–34. doi:10.1016/j.jns.2015.10.007.

[289] S. Lee, A. Raw, L. Yu, R. Lionberger, N. Ya, D. Verthelyi, A. Rosenberg, S. Kozlowski, K. Webber, J. Woodcock, Scientific considerations in the review and approval of generic enoxaparin in the United States., Nat. Biotechnol. 31 (2013) 220–6. doi:10.1038/nbt.2528.

[290] S.A. Berkowitz, J.R. Engen, J.R. Mazzeo, G.B. Jones, Analytical tools for characterizing biopharmaceuticals and the implications for biosimilars., Nat. Rev. Drug Discov. 11 (2012) 527–40. doi:10.1038/nrd3746.

[291] A. Bufe, E. Von Mutius, O. Holst, C. Braun-Fahrländer, D. Nowak, J. Riedler, Stable dust extract for allergy protection. EP 1637147 B1, EP 1637147 B1, 2004.

[292] M. Guerrini, A. Bisio, G. Torri, Combined quantitative (1)H and (13)C nuclear magnetic resonance spectroscopy for characterization of heparin preparations., Semin. Thromb. Hemost. 27 (2001) 473–82. doi:10.1055/s-2001-17958.

[293] W.L. Chuang, H. McAllister, L. Rabenstein, Chromatographic methods for product-profile analysis and isolation of oligosaccharides produced by heparinase-catalyzed depolymerization of heparin., J. Chromatogr. A. 932 (2001) 65–74. http://www.ncbi.nlm.nih.gov/pubmed/11695869 (accessed November 17, 2016).

[294] H. Helmby, Human helminth therapy to treat inflammatory disorders - where do we stand?, BMC Immunol. 16 (2015) 12. doi:10.1186/s12865-015-0074-3.

[295] J.R. Feary, A.J. Venn, K. Mortimer, A.P. Brown, D. Hooi, F.H. Falcone, D.I. Pritchard, J.R. Britton, Experimental hookworm infection: A randomized placebo-controlled trial in asthma, Clin. Exp. Allergy. 40 (2010) 299–306. doi:10.1111/j.1365-2222.2009.03433.x.

[296] A.S. Navarro, D. Pickering, I.B. Ferreira, P.R. Giacomin, Title : Hookworm recombinant protein promotes regulatory T cell responses that suppress experimental asthma, 143 (2012) 1–15. doi:10.1126/scitranslmed.aaf8807.

[297] K.G. Nicholson, J.M. Wood, M. Zambon, Influenza., Lancet (London, England). 362 (2003) 1733–45. doi:10.1016/S0140-6736(03)14854-4.