3-26-2021 9:00 AM

A Workflow to Analyze ETHcD Mass Spectrometry Data for Studying HIV gp120 Glycosylation

Yingxue Sun, *The University of Western Ontario*

Supervisor: Creuzenet, C., *The University of Western Ontario*
Co-Supervisor: Arts, E., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Microbiology and Immunology

Abstract

The great heterogeneity of HIV populations and richness of surface glycan clouds makes it difficult to locate a conserved and exposed protein epitope as an effective vaccine target. However, more than 80% new infections result from single transmitted founder (T/F) viruses. We set out to design a workflow to study the traits of T/Fs that allow for their superior infectivity, specifically, the glycosylation patterns of gp120, a subunit of HIV envelope protein responsible for binding to host cell receptors. Our main research methods include Western blot and mass spectrometry. Our current understanding of the mass spectrometry data indicates that our T/F and chronic HIV strains have differential distributions of glycan density at several key N-sites throughout the gp120 peptide backbones, which may be related to the differential transmission fitness of the two strains and potentially used as novel glycopeptide-based HIV vaccine targets.

Keywords

HIV (human immunodeficiency virus), HIV sexual transmission, transmission fitness, transmitted founder (T/F) virus, HIV envelope protein (Env), gp120, glycosylation, glycobiology, mass spectrometry, post-translational modification, vaccine strategy.

Summary for Lay Audience

One of the greatest challenges in HIV vaccine development is that the viral populations are highly diverse, making it difficult to find a universal vaccine target that works on all viruses. Despite the large genetic diversity, more than 80% of new HIV infections result from the transmission of a single virus, known as a transmitted founder (T/F) virus. This indicates that usually only one T/F out of the pool of viruses from the HIV donor is able to establish a stable infection in a new recipient. Understanding the unique features of T/Fs will provide novel strategies for vaccines and ultimately prevent the spread of HIV worldwide.

We set out to design a workflow to study one of the major factors believed to give T/Fs a selective advantage during transmission: gp120 glycosylation. Gp120 is a protein on the surface of HIV particles and initiates the infection of a human host cell. Glycosylation is a network of sugar chains (or glycans) attached to the protein backbone at specific points known as N-linked sites. As HIV undergoes frequent genetic mutations, viruses evolve to have different numbers and locations of N-linked sites and different types of sugar chains. The goal of my project is to compare the gp120 glycosylation profiles of a T/F strain and a strain derived from chronic stage of untreated HIV-1 infection. The major tool we use is mass spectrometry, which breaks down gp120 into small fragments of peptides, sugars, and glycopeptides. This allows us to identify the types of the glycans present in the sample and at which N-linked sites they are attached. We then generate a matrix of N-linked sites and types of glycans for both strains. In this thesis, our findings show that the two strains have differential distributions of glycan density throughout the gp120 peptide and that they express distinct compositions of glycans.

This project is a proof-of-principle study that provides tools for larger-scale studies to include more strains of T/F and chronic HIV viruses. Ultimately, this workflow will help unveil the key glycopeptide signatures accountable for HIV transmission that can be used as novel vaccine candidates.

Co-Authorship Statement

This thesis does not include published or non-published data of another candidate. However, training and help for certain experimental methods were provided by Adam Meadows, Dr. Najwa Zebian, and Dr. Katja Klein, along with all other members in the Creuzenet and Arts labs.

Acknowledgments

To begin, my biggest thanks go to my supervisors and role models, Dr. Carole Creuzenet and Dr. Eric Arts. Without your kind guidance and continuous support, especially during the COVID-19 crisis, I could not have made the progress so far. Thank you for giving me the freedom to be creative and design new methodologies while helping me stay on track to meet my goals. It has been a great pleasure and honor to have this great learning experience as your trainee. I am also tremendously grateful for my committee advisors Dr. Jamie Mann and Dr. Art Poon. Thank you so much for being generous with your time, expertise, and resources to help improve my research skills throughout this project.

Continuing on, I'd like to thank all the members in both Creuzenet and Arts labs, especially Dr. Najwa Zebian and Adam Meadows, whose work laid solid scientific foundations for my project. Thank you all for offering technical help and making our lab a cheerful big family. I extend my sincere gratitude to my good friend and colleague from Robarts Institute, Wenyao Xia, who kindly helped me get better at troubleshooting programming problems. Pursuing science can be tough and even frustrating at times, but it is in the friendship with you all, I never felt lonely in this journey. Of course, I could not have met all of these incredible mentors and peers if not for the MNI department, Western University, and the funding agencies. I thank all the PIs, staff, and fellow trainees for their hard work that built our community collectively.

Finally, I give special thanks to my family. Mom and dad, thank you for sparking my interest in science and research, being my support system, and most importantly, inspiring me to pursue a meaningful life. Jimmie, thank you for letting me borrow your gaming computer for my data analysis and for saving the snacks from my stress eating.

I'm truly blessed for all the individuals who contributed to help me become a better scientist through my MSc study. I could not have done this without you all.

Table of Contents

List of Tables

List of Figures

List of Appendices

1 Introduction

List of Abbreviations:

HIV: Human Immunodeficiency Virus

DC: dendritic cells

LC: Langerhans cells

AIDS: acquired immunodeficiency syndrome

ART: antiretroviral therapy

T/F: transmitted founder virus

Env: envelope proteins

STD: sexually transmitted disease

PNGS: potential N-linked glycosylation sites

APC: antigen presenting cells

PRR: pattern recognition receptors

MBL: mannose binding lectin

GNA: galanthus nivalis agglutinin

FT: Flowthrough

MS: mass spectrometry

ETHcD: electron-transfer/higher-energy collision dissociation (ETHcD)

ETD: electron-transfer dissociation

HCD: higher-energy collision dissociation

NSD: N-Site Distrubution, the proportion of glycans at a specific N-site out of the total number of glycans found in the entire dataset.

RA: Relative Abundance (RA), the proportion of glycans of a specific glycoform out of the total number of glycans found in the entire dataset.

HSP: Homologous Site Pair, the pairing of N-sites in the B4 and Q0 sequence based on homology, rather than the order of appearance along the protein sequence.

The human immunodeficiency virus (HIV) is a retrovirus that infects human cells expressing the CD4 receptor and either the CXCR4 or CCR5 co-receptor[1]. While the most permissive host cell type to HIV is the T helper cell family, some subpopulations of dendritic cells (DCs), Langerhans cells (LCs), and macrophages are also susceptible reservoirs[1]. HIV can be transmitted through various sexual activities, sharing needles and syringes, or from mother to child during pregnancy, birth, and/or breastfeeding[2]. This study will discuss transmission by unprotected sex, the most common cause of HIV, where the risk of infection is 0.65-1.7% per anal intercourse and 0.03-0.5% per heterosexual intercourse[3]. Following transmission, there is a 10-to-12-day period known as the eclipse phase, where HIV cannot be detected by any diagnostic test as the HIV RNA is yet to reach detectable levels[4]. As HIV continues to replicate, the newly infected individuals start to experience a few weeks of flu-like symptoms known as the acute phase, where the virus is rapidly undergoing replication. Infected individuals then gradually enter into the chronic phase of infection which can last for around 10 years and do not exhibit apparent symptoms[5]. If left untreated, patients with chronic HIV infection will progress to acquired immunodeficiency syndrome (AIDS), a condition in which the risks of opportunistic infections and cancers become very high due to the depletion of CD4 T cells[1,5]. If treated with antiretroviral therapy (ART), the patients can remain in the manageable asymptomatic condition for decades[5,6]. During ART, viral replication is suppressed. Despite some viral blips, the occasional transient increase in viral load in some patients, the viruses are kept at undetectable levels and it is therefore untransmittable to HIV-negative sexual partners[6,7]. It is estimated that 38 million people are living with HIV globally and 25.4 million people now have access to ART globally, representing approximately 67% of the total cases. While over 80% of patients in Western-Central Europe and North America are receiving ART, only around 40% of patients from Eastern Europe, the Middle East, North Africa, and Central Asia have access to effective ART[8].

HIV is divided into two large subspecies, HIV types 1 and 2. While HIV-1 accounts for 95% of the global epidemic, HIV-2 is localized mostly in West Africa and associated with slower disease progression and lower transmission rate. HIV-1 is categorized into groups M, O, and N, each resulting from a separate zoonotic transmission[9,10]. Group M is responsible for 90% of the total HIV-1 cases around the world and is further divided into many subtypes or clades based on their geographical distribution, including subtype A through L[10]. Notably, subtype B is common to North America and Europe and is the most studied. Subtypes A and D are prevalent mostly in Africa, while C and E dominates areas along the Indian Ocean coastline and South East Asia[9,10]. Since each HIV particle has two copies of single-stranded RNA, it is possible to generate inter-clade recombinant strains during viral production and assembly in hosts with infections of more than one HIV strain, or superinfections[10]. The recombinants are known as circulating recombinant forms (CRFs), and exhibit with faster disease progression and are more likely to develop drug resistance to ART[10,11]. The emergence of CRFs necessitates effective vaccines and novel medication strategies be developed.

1.1 Transmission cycle and major bottlenecks

To broaden the current spectrum of HIV treatment options, it is crucial to understand viral transmission from an infected donor to an HIV-negative recipient. Starting from the donor's blood, during the chronic stage of untreated infection, the virus is abundant and genetically diverse, despite the presence of circulating anti-HIV antibodies[12]. To be transmitted to the recipient via sex, it is believed the virus must be present within the donor's genital tract by replicating locally and/or by migrating there from the blood[12]. The donor's genital tract seems to be a selective environment due to the presence of neutralizing antibodies and opsonizing lectins and is often populated by virions amplified from only several distinct variants[12]. Some studies suggest that the expansion of these clonal populations is likely facilitated by inflammation in the donor's genital tract resulting from micro-tears during sex or coinfection with other sexually transmitted diseases (STDs)[12-14]. Local inflammation recruits more immune cells that are permissive to HIV infection and replication[12-14]. As a result, the viral load is increased in the inflamed genital tract of the untreated donor, posing a stronger transmission potential.

The viruses are shed into secretory fluids such as mucus or semen which come into contact with the next host through sexual intercourse[12,13]. Once within the genital tract of the recipient, the viruses must cross the mucosal and epithelial layers in order to access their target CD4+/CXCR4+ and CD4+/CCR5+ cells[12,15-17]. This step is believed to be the most stringent genetic bottleneck in the HIV transmission cycle: so stringent that, in 60-80% cases, only a single virion in a pool of millions of genetically distinct variants is able to establish a stable infection. This successful variant is known as the Transmitted Founder (T/F)[12,15-17]. As the T/F undergoes initial rounds of replication in the new host, the viral population is highly homogeneous[12]. Over time, mutations accumulate in the viral genome and the viral population becomes extremely diverse as the patient progresses from the acute phase to the chronic phase of infection[12,15].

1.2 HIV envelope protein (Env) and gp120 glycosylation

T/F viruses have been of significant interest since their discovery and many groups have studied which phenotypic features can bestow T/F virus' superior transmission fitness. One of such features believed to bestow T/F with increased transmission fitness is their envelope protein (Env), particularly the gp120 subunit[16,18]. Envs are spike-like structures extending out from the spherical viral surface. Each functional Env is a trimer that consists of 3 heterodimers of gp120 and gp41. The gp120 outer portion sits on the transmembrane gp41 partner and is responsible for recognizing and binding the host cell receptors, shown in Fig 1a. Gp41 serves to facilitate the fusion of viral and cellular membranes once gp120 is bound, releasing the viral contents into the cell[19-21]. Gp120 and gp41 are synthesized as a single precursor known as gp160, which becomes trimerized, glycosylated in the ER and the Golgi, and then cleaved into gp41 and gp120 subunits by the host enzyme furin[21-23]. The gp120 subunit differs significantly between variants in both protein sequence and glycosylation patterns[25]. Glycosylation is a type of post-translational modification where oligosaccharide chains, or glycans, are attached onto the protein backbone at specific sites by enzymes known as glycosyltransferases[25]. Glycosylation of gp120 is extremely important for viral survival as the glycan shield protects the inner protein core from antibody mediated detection. In fact, approximately 50% of gp120 mass is comprised of carbohydrates[26-30]. Nonetheless, in an evolutionary

arms race, the host may develop broadly neutralizing antibodies, such as 2G12 and PGT121, which are able to target gp120 glycan epitopes, shown in Fig. 1b[26-30]. However, the neutralizing capacities of these antibodies largely vary from patient to patient[28-29].

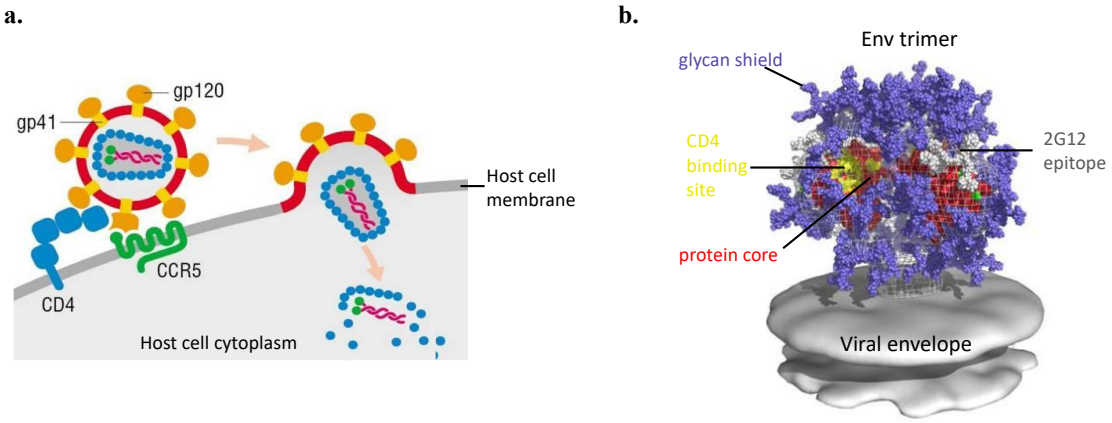


Figure 1. Function and structure of gp120 in HIV transmission.

(a). HIV gp120 directly interacts with host cell's CD4 receptor and CCR5 (or CXCR4) coreceptor. (b). The gp120 protein (red) is heavily glycosylated to form the glycan shield (blue). Certain antibodies, such as 2G12, could bind gp120 by glycan-based epitopes[30].

Over-glycosylation can be highly detrimental to the virus as it makes the virus susceptible to host lectins, a family of opsonizing immune proteins present in the mucosa that bind carbohydrates on pathogens[32-34]. Therefore, viruses must maintain a well-balanced as well as dynamic glycosylation profile in response to different threats at different stages of its life cycle, discussed in detail in the following section.

1.3 Gp120 glycosylation in T/F and chronic viruses

Since viral protein synthesis relies on the host cell machinery, gp120 may undergo both types of glycosylation pathways found in the human system: N- and O-linked[25]. However, N-linked glycosylation has been shown to be much more abundant and have far greater impact on HIV fitness and infectivity than O-glycosylation[34-37]. N-glycosylation may only occur at specific motifs (N-X-S/T, where X can be any amino acid but proline) on the polypeptide sequence known as potential N-linked glycosylation sites (PNGS). Not all PNGS are glycosylated as their positions in the folded gp120 may

limit their accessibility to glycosyltransferases[25]. The number of PNGS in the gp120 sequence varies between 18 to 33 among different subtypes with a median of 25[38,39]. Studies have found that T/F gp120 typically has fewer PNGS than its chronic counterpart[18,40,41]. As the viral protein sequences mutate and evolve over time, more PNGS could be created in the chronic viruses, allowing for more opportunities to build up the glycan shield in response to the neutralizing antibodies gradually produced in the host[18,40,41]. Indeed, for the T/F viruses, over-glycosylation can not only increase the risk of mucosal lectin detection but also physically hinder gp120 binding to target cell receptors. The protective function of the glycan shield against the neutralizing antibodies is also less critical at the transmission stage since no antibodies are yet produced in the new host. On the other hand, T/F viruses must maintain some glycans because they enhance viral attachment to the migratory antigen presenting cells (APCs) via the cell membrane-bound lectin receptors, increasing the probability of infection once the virion-carrying APCs cross the mucosal barrier and present the viruses to the T cells[42-49]. The hijacking of APCs by HIV and the relevant lectin receptors are discussed in detail in section 1.7.

Besides the reduced number of PNGS, most glycans in T/F gp120 are thought to be of the high-mannose type, while the chronic viruses contain a greater diversity of complex type glycans[18,23,28,29,40,50]. The glycosylation process depends upon the expression of enzymes in the host cell type which can work together to create a variety of glycoforms[25]. Glycosylation of Env begins in the endoplasmic reticulum (ER), where the precursor glycan, which contains 2 N-acetylglucosamine (GlcNAc), 9 mannose, and 3 glucose residues, is first attached to the nascent peptide by an enzyme known as oligosaccharyltransferase (OST) at accessible N-X-S/T motifs[25]. The precursor glycan then undergoes initial steps of modifications, such as trimming of glucose residues by α-glucosidases I and II and mannose residues by α-mannosidase I, resulting in various high-mannose forms with 8 or 9 mannose residues[25]. Then, the glycoprotein is transported from ER to the *medial*- and *trans*-Golgi for further glycan processing, such as addition of GlcNAc residues by N-acetylglucosaminyltransferases I and II, galactose residues by galactosyltransferases, and/or sialic acid residues by sialyltransferases[25]. Glycan maturation in the Golgi compartments gives rise to a great variety of hybrid and complex

glycoforms[25]. The predominance of high-mannose glycans and the minority of mature forms in T/F gp120 glycosylation profile indicate reduced Golgi processing during gp120 protein synthesis. However, late Golgi processing cannot be completely excluded as the cleavage of gp160 by furin, which is required for proper viral assembly, occurs in the *trans* Golgi network[21,22]. The high-mannose phenotype in T/F gp120 is likely a result of physical inaccessibility of the Golgi-resident glycan-processing enzymes due to the steric hindrance imposed by the folded protein structure[23,40].

1.4 Coreceptor tropism and transmission fitness

Another T/F-distinct feature is that over 90% T/F are found to be R5-tropic, regardless of strain and subtype[51-53]. R5-tropic viruses preferentially use the CCR5 coreceptor, whereas X4-tropic viruses use CXCR4, in addition to the universally required CD4 receptor during infection. Furthermore, between the two known conformations of CCR5, one is more sensitive to maraviroc, a CCR5 antagonist, and is preferentially used by T/Fs[51,53]. R5/X4 tropism is largely determined by the V3 region of gp120, which shapes the interaction surface with coreceptors[55-58]. It was found that V3 contains an exposed area formed by amino acid residues 11 and 24 or 25, which is positively charged in X4-tropic viruses and negatively charged or neutral in R5-tropic viruses[55].

Pre-incubation with HIV-free seminal plasma can protect CD4+ T cells from later challenges with both R5- and X4-tropic viruses through the downregulation of CD4 surface expression[57]. Meanwhile, the seminal plasma also upregulates surface expression of CCR5, partially counteracting the protective effect against R5-tropic viruses[60]. It has also been shown that some migratory immune cells, such as dendritic cells (DCs) and macrophages, can robustly increase the transmission efficiency of HIV by100- to 10,000-fold by trafficking the virions from the mucosa to T cells for antigen presentation[56]. DCs and macrophages have been shown to almost exclusively select R5-tropic viruses[61,62].

1.5 Physical barriers against HIV

Breaking through the recipient’s genital mucosal epithelial barrier is the key to T/F success in all forms of sexual transmission. The epithelia of foreskin, glans penis, labia,

vagina, ectocervix, and fossa navicularis are made of multilayered stratified squamous epithelial cells[63,64]. The urethra, anal canal, rectum, and endocervix are only lined by a single layer of columnar epithelial cells, which makes these areas weak points for viral infiltration[63,64]. The per-exposure risk of acquiring HIV for the insertive party and receptive party is 0.11% and 1.4% in anal sex, respectively, and 0.04% and 0.08% in vaginal sex, respectively[63]. This is consistent with the stringency of the epithelial barriers in the anorectal and vaginocervical canals[66]. Three models of HIV pathogenesis from genital inoculation to systemic infection have been proposed for the male-to-female mode of transmission[67]. The first model is based on the assumption of the vaginal mucosa being the most stringent bottleneck and states that only a single virus can cross the epithelium and then establish a systemic infection in the recipient[67]. The second model also suggests that a single virus can penetrate the mucosa, but it first replicates and evolves in local tissues, producing several distinct variants, one of which may then be disseminated into the circulatory and the lymphatic systems[67]. The third model, like the second model, also describes the localized foci of infection in the mucosa of the vagina and cervix. However, rather than evolving from a single virus, these foci are results of several viruses from the inoculum[67]. Both model 2 and model 3 consider an additional bottleneck that selects only one of the localized viral clones to establish a systemic infection. The foci of infection in the genital mucosa are formed by viral replication in the local CD4+ T cells, infiltrating CD4+ T cells that are recruited in response to the inflammatory environment, and dendritic cells (DCs)[67,68,69]. These infected cells can then disseminate the virus systemically. HIV interactions with T cells and DCs will be discussed in detail in section 1.7.

Above the epithelial cell barrier exists the mucosa, home of many commensal microbes. *Lactobacilli* are the dominant species in the healthy female vagina and can secrete the virucidal substances H_2O_2 and lactic acid[70,71]. Other bacteria are not shown to have direct effects on HIV, however their role in maintaining a homeostatic inflammatory environment is crucial in immune defense. It has been shown that individuals with genital tract inflammation are more prone to HIV infection via sex[71]. This is especially true for individuals co-infected with other sexually transmitted diseases (STDs)[71,76]. The gut microbiota profile has been used to predict one's risk of HIV infection and

responsiveness to pre-exposure prophylaxis strategies[73,74]. Fecal transplant has shown promising synergistic effects with ART in patients coinfected with HIV and *Clostridium difficile*[75,76].

More T/F studies have focused on the acquisition of HIV via vaginal and anal routes rather than the penile route. An observation that circumcision can reduce HIV transmission rates by over 60% sparked curiosity into the mechanisms of female-to-male (FTM) transmission through the foreskin, a bilayer structure where the inner face lies on the glans and the outer face folds back-to-back on the inner face[77,78]. Initial studies identified the inner foreskin as a risk factor as its thinner keratin layer can increase target cell exposure to HIV; however, more recent studies found no difference in keratin thickness between circumcised and uncircumcised men[79,80]. Prodger and colleagues have shown that tissue resident immune cells can cluster in foci, providing an enriched environment for potential infection and robust replication[81]. The increased risk of HIV transmission in uncircumcised men is likely due to larger skin areas containing greater numbers of such foci. Further, foreskin is shown to promote local inflammation, compromising epithelial barrier integrity and recruiting HIV susceptible T cells[82]. Post circumcision, the level of pro-inflammatory IL-8 is shown to gradually decline over 2 years or longer[83]. It is also shown that the fold between the inner foreskin and the glans creates an anaerobic environment, which leads to a higher proportion of gram-negative bacteria in the microbiome of uncircumcised men[83]. Specifically, *Prevotella spp*., an anaerobic bacterium associated with bacterial vaginosis and elevated IL-8 levels, is shown to constitute 20% of the total foreskin microbiomes in 87% of uncircumcised Ugandan men[83]. In summary, circumcision reduces the rate of FTM HIV transmission likely through decreasing the number of immune cells foci and the level of pro-inflammatory cytokines.

1.6 HIV interactions with hosts' proteins

At the site of inoculation, the body fluids of the donor and recipient come in contact, such as semen and mucus. A number of biomolecules in these fluids can directly interact with HIV virions that either enhance or reduce viral infectivity. Semen is believed to enhance

infectivity in two ways[84]. First, semen contains a variety of signaling molecules including cytokines, Transforming Growth Factor β (TGFβ), and Prostaglandin E2 (PGE2), which mediate cervical tissue inflammation and increase the epithelial permeability to viral particles[84,85]. Second, semen-derived enhancers of virus infection (SEVI), a type of protein aggregate fibril in semen, is shown to promote viral attachment to host cells via its cationic interactions with negatively charged molecules, such as heparan sulphates, on cell membranes. The semenogelin protein has similar effects as SEVI and the semen of semenogelin-deficient donors lacks this enhancement of HIV infectivity[84,85]. Unlike semen, the vaginal and anal mucosal fluids are much more deleterious environments to viruses as they contain various types of innate immune proteins. Lectins are a large class of pattern recognition receptors (PRRs), including C1q and mannose binding lectin (MBL), that specialize in recognizing pathogens carrying carbohydrates, e.g., gp120 glycans on HIV envelope[87,88]. Research on HIV-lectin interactions started with an observation that patients who have variant alleles coding for reduced levels of MBL are subject to increased susceptibility to HIV transmission[89]. Lectins can be secreted into the mucus or expressed on cell plasma membrane. The soluble lectins can reduce viral infectivity in at least two ways. First, lectins can act as physical traps that bind to HIV Env, which retains the viruses in the mucus and prevents further entry into the submucosa[87,90]. Second is opsonization. Each lectin has six sugar binding heads and can form an immune complex with two zymogens. For example, MBL is partnered by the MBL-associated serine proteases (MASP-1 and MASP-2). When two or more lectin heads are bound to sugars from pathogens, the zymogens are auto-cleaved, resulting in the production of the active opsonin C3b and initiation of the complement pathway, where the pathogens are engulfed by phagocytes[32,90,91]. The third aspect of lectin actions involves killing the invader by pore formation on its surface, yet this microbicidal effect has only been confirmed for bacterial and fungal pathogens, but not viruses[91-94]. Interestingly, beyond human lectins, studies have found a wide variety of plant-, cyanobacteria-, worm-, and even chemical-derived lectins that are highly potent in blocking HIV[95]. A few examples are galanthus nivalis agglutinin (GNA), griffithsin, cyanovirin-N (CVN), etc. The therapeutic application of these lectins is another fascinating field but will not be discussed in detail in this thesis. Of course, some lectins

are not able to block HIV infections. A soluble lectin, galectin-1, has been demonstrated to assist HIV-1 infection *in vitro* and *ex vivo* by augmenting viral adhesion to susceptible immune cells. Paradoxically, studies have also shown that complete opsonization of HIV particles can increase infectivity by promoting viral internalization by DCs, which can deliver viruses to T cells in a Trojan Horse fashion during antigen presentation[108]. This Trojan Horse model is discussed in detail in the following section.

1.7 HIV interactions with hosts' cells

HIV's primary target host cell type is CD4+ T helper cells, including naïve and memory T cells[97,98]. The memory T cells contain subgroups such as the central and effector memory T cells, which populate in the secondary lymphoid tissues and peripheral tissues, respectively[99]. As mentioned previously in section 1.2, the first steps of HIV infection require viral attachment and entry into the target cell through the HIV envelope protein's interactions with the cell surface CD4 receptor and one of the CCR5 or CXCR4 coreceptors[1,17,19,58]. CCR5 is expressed in the effector memory T cells in sites such as the vaginal and intestinal mucosa, while CXCR4 is predominantly found in the naïve T cells[100,101]. As introduced in section 1.4, the CCR5 coreceptor is preferentially used by over 90% of T/F viruses, while the use of CXCR4 is typically developed later on in a subpopulation of HIV viruses in chronically infected patients[54-56,102-104]. In summary, the mucosa-resident CD4+ CCR5+ effector memory T cells are the most important target for HIV sexual transmission.

Env is comprised of three heterodimers of gp120 and gp41, forming the crown and the transmembrane anchor, respectively[24-26]. Gp120 has five conserved regions (C1-C5) and five variable regions (V1-V5)[105]. The core of the trimeric gp120 crown is formed by the conserved regions and contains the CD4 binding site[106]. The variable regions form the surface loops[102]. The binding site for CCR5/CXCR4 is formed by two domains of gp120 known as the bridging sheet and V3 loop[106]. However, the coreceptor binding site is only revealed upon the conformational changes caused by the binding between gp120 and CD4[106]. When gp120 binds both the receptor and the coreceptor, gp41 can then be inserted into the cell membrane and mediate the fusion of viral and cellular

membranes[107]. The viral contents are released into the host cell, such as the viral genome, reverse transcriptase, and integrase[107]. These materials will be used to generate and permanently insert the DNA copy of the viral genome into the host genome[107].

While the many types of soluble lectins are known to be inhibitory to viruses, many cell-bound lectins are notorious for their role in enhancing HIV infectivity[45-53]. Some notable examples of such lectins include Dendritic Cell-Specific Intercellular adhesion molecule-3-Grabbing Non-integrin (DC-SIGN), macrophage mannose receptor (MMR), sialic acid-binding Ig-like lectin 1 (Siglec-1), and dendritic cell immunoreceptor (DCIR)[45-53]. These lectins are expressed on DCs and macrophages, which are migratory in nature and members of the antigen presenting cells (APCs)[45-53]. APCs are cells that capture foreign pathogens and travel to the lymph nodes to present the foreign antigens to the T cells for potential activation of the adaptive immune response[108]. APCs are recruited to sites of micro-tears of the genital mucosa created from sexual activities, where HIV viruses can enter the tissue and associate with the APCs via the binding of the cellular lectin and the viral gp120 glycans[45-53]. As the APCs migrate to the T cells for antigen presentation, HIV viruses can hijack this pathway and be passed conveniently from the APCs to the T cells *in trans* at the infectious synapse[109]. Moreover, these APCs themselves are also susceptible to HIV infection and replication since they also express CD4 and CXCR4/CCR5 coreceptors[1]. Viral clones may be produced *de novo* in APCs as soon as 24-72 hours after challenge and can then infect target T cells in *cis*[46,109,110].

Langerhans cells (LCs) are members of the DC family found in the epidermis and stratified mucosal epithelia[1, 51]. In addition to CD4, CCR5, CXCR4, and many lectins found expressed on DCs, they uniquely and richly express a C-type lectin known as langerin[51,111]. HIV particles bound by Langerin are internalized into Birbeck granules in LCs and degraded[111]. However, at high viral concentration, Langerin receptors can be saturated, enabling HIV binding to CD4 and CCR5[111]. Infected LCs can then transfer HIV to T cells primarily in a *cis* fashion[112].

Although this cell-assisted method is the most efficient way of HIV entry, it is important to recognize that HIV may also use two other less efficient routes to traverse the

epithelial cell layer: paracellular diffusion (movement through the interstitial space) and transcytosis (movement through the cytoplasm)[110]. The cell-cell transmission of HIV is accepted as 100- to 10,000-fold more efficient than fluid-phase viruses[59].

Paracellular diffusion of HIV particles (120 nm in diameter) may seem impossible as the intact epithelial layer is usually impermeable to molecules larger than 10 nm[110]. However, epithelial cell junctions may be loosened or absent in inflamed or physically damaged conditions, which are common in sexual activities[111]. Diffusive viral penetration is shown to occur in both stratified squamous and monolayer columnar epithelial barriers in human vaginal and cervical explants and macaque models[112,113]. The percolation efficiency through the epithelium is best correlated with the initial viral load in the inoculum and less so with the mere barrier thickness[114].

In contrast to passive diffusion, transcytosis is a receptor-mediated, targeted process where cargos are contained in vesicles and trafficked across a cell membrane[115]. Although epithelial cells lack the canonical CD4 receptor for HIV's Env to bind, endocytosis can still be achieved with alternative receptors, including GalCer (galactosylceramide, a glycosphingolipid), HSPGs (heparan sulphate proteoglycans), scavenger receptor gp340, ICAMs (Intercellular Adhesion Molecules), and the Fc neonatal receptor (FcRn)[113,116].

1.8 Summary, aims, and hypothesis

In summary, the goal of this study is to develop a workflow to investigate the traits that allow for T/F virus' superior infectivity by studying the glycosylation patterns of gp120 in a T/F and a chronic strain of HIV, namely B4 and Q0, respectively. Specifically, we aim to use mass spectrometry to determine both the locational distribution of glycans along the gp120 peptide chain and the identities of glycoforms at each N-linked site. We hypothesize that the gp120 purified from B4 and Q0 have differential glycan distribution as well as glycoform composition. This work will provide a proof-of-principle example for future projects unveiling the key glycopeptide signatures accountable for HIV transmission as novel targets vaccine and prophylaxis strategies.

2 Materials and Methods

2.1 Gp120 purification from HIV clones

Transmitted Founder virus and chronic stage virus were used in these studies to generate B4 and Q0. The viruses were propagated in the Arts lab at the University of Western Ontario. The B4 and Q0 vectors were initially cloned by inserting the *env* genes (B4 *env* gene obtained from Center for HIV-1/AIDS Vaccine Immunology and Q0 *env* gene from Case Western Reserve University) into the pREC_nfl_NL4-3_Δenv/URA3 HIV backbone using yeast-based recombination to produce the pREC_nfl_HIV-1 plasmid[117]. The resulting pREC_nfl_HIV-1 plasmid and a complementing vector, pCMV_cplt, which contains the unique 5′ element, repeat of the LTR, and the Ψ packaging signal, were co-transfected into 293T cells using the Effectene lipid system (Qiagen) to produce infectious virions[117]. The viruses were then propagated in the CD4+CCR5+ U87 cell line (NIH AIDS Reagent Program). Cells were cultured using DMEM medium (Sigma-Aldrich) supplemented with 10% fetal bovine serum, 1% Penicillin/Streptomycin (Sigma-Aldrich), 300 μg/ml G418 (Thermofisher Scientific), and 1 μg/ml puromycin (Thermofisher Scientific) for a maximum of 10 passages. Viruses were harvested every 3 days until day 14 by pooling the tissue culture medium and stored at -80 ^{o} C until at least 1 L total volume was reached for each biological replicate.

After thawing, the medium containing either B4 or Q0 virus was first spun at 3500x g using the Beckman Coulter Allegra 64R centrifuge for 5 minutes to pellet cellular debris. Then the supernatant containing proteins and virus was spun using ultracentrifugation to pellet the viruses (Beckman Coulter SW 32 Ti Rotor, 100 000 g, 2 hours, 4 ^{o} C). Ultracentrifugation supernatant was transferred to 50-ml falcon tubes and stored at -80 ^{o} C for future use (described in section 2.2). Each pellet was resuspended in 500 μl of 100 μM ammonium bicarbonate at pH 8.0 and transferred to 15-ml falcon tubes from the ultracentrifugation tubes. Benzonase (Millipore Sigma, ≥250 units/μl) and $MgCl_2$ were added to the resuspended virus to final concentrations of 2.5 unit/ml and 2 mM, respectively, to degrade free DNA or RNA. Empigen BB, a detergent, (Sigma-Aldrich, 30% stock w/v) was added to a final concentration of 0.25% to lyse the viruses and

release gp120 from the viral envelope. The mixture was vortexed and incubated at room temperature for 2 hours without agitation. NaCl was added to a final concentration of 650 mM to stop Empigen BB's detergent activity. The lysate was centrifuged at 64400x g for 15 minutes, and the supernatant containing free gp120 is filtered using 0.8 μm filters to remove the viral debris.

Gp120 was purified using a column containing 1 ml agarose-conjugated GNA lectin (Sigma-Aldrich, binding capacity 5-10 mg/mL). Ten column volumes (CVs) of equilibration buffer (20 mM Tris-HCl pH 7.5, 650 mM NaCl, 0.25% Empigen BB) was first passed through the column. Filtered viral lysate was then allowed to pass through the column by gravity. The column was washed with 10 CVs of 20 mM Tris-HCl pH 7.5, 650 mM NaCl to remove excess Empigen BB, followed by 10 CVs of 20 mM Tris-HCl pH 7.5, 1 M NaCl to remove nonspecifically bound proteins, and 10 CVs of 20 mM Tris-HCl pH 7.5, 150 mM NaCl to remove excess salts. Gp120 is then competitively eluted from the GNA lectin using 10 CVs of 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 M methyl-α-D-mannopyranoside. Elution was dialyzed overnight at 4 ^{o}C in dialysis tubing (Fisher Scientific, 12-14 kDa MWCO) in 50 mM ammonium bicarbonate at pH 7.5 with gentle magnetic stirring. Dialyzed sample was frozen at -80 ^{o}C for 24 hours, lyophilized at -50 ^{o}C and 0.3 mbar, and resuspended using 75 μl of 50 mM ammonium bicarbonate at pH 7.5 and stored in -20 ^{o}C.

2.2 Concentration of shed gp120 in culture supernatant

As described in section 2.1, the tissue culture medium containing viral particles was harvested and the viruses were pelleted using ultracentrifugation. The supernatant containing shed gp120 was kept for two virion batches: B4-5 and Q0-6. The supernatant was loaded onto centrifugal filter units with 100 kDa MWCO (Ultracel-100 regenerated cellulose membrane with 15 ml sample volume, Sigma-Aldrich) and spun at 3500x g using the Beckman Coulter Allegra 64R centrifuge at 20-minute intervals. The flowthrough containing irrelevant proteins, such as albumin from the FBS, was discarded. Gp120 was concentrated in the liquid above the filter and was collected.

2.3 Visualization of gp120 using Western blot

Samples were mixed with 4x SDS loading buffer (0.625 M Tris, 2% SDS, 2% β-mercaptoethanol, 10% glycerol, 0.002% bromophenol blue, pH 6.8) in 3:1 ratio. The mixture is denatured at 100 ^{o}C for 5 minutes. SDS-PAGE was run using 0.75 mm 10% polyacrylamide gels (made in lab) at 150mV for 65 minutes (Bio-Rad mini gel system) in running buffer (0.025 M Tris, 0.192 M glycine, 0.1% SDS, pH 8.3). Protein bands are then visualized using either Western blot or Coomassie staining.

For Western blot, protein bands from the gel were transferred to a Nitrocellulose membrane (Bio-Rad) for 45 min at 180 mA with cooling in transfer buffer (25 mM Tris, 192 mM glycine, 10% methanol, pH 8.3). The membrane was immediately rinsed with milliQ water and stained with Ponceau red (0.1% Ponceau S (Sigma-Aldrich) in 1% acetic acid) for a few minutes. The membrane was then de-stained using 1x PBS buffer until protein bands could be distinguished from the background. The membrane was scanned and then blocked in 10% skimmed milk in 1x PBS buffer with gentle agitation for 1 hour at room temperature. The membrane was washed with 1x PBS with 0.02% Tween-20 for 5 minutes twice and with 1x PBS for 5 minutes once. Primary antibody was applied to membrane and incubated for 1 hour at room temperature. The membrane was washed as described above and incubated at room temperature with either fluorophore- or HRP-conjugated secondary antibody for 35 minutes or 1 hour, respectively. The membrane was washed again before imaging.

The primary antibody used, B13, was produced from mouse hybridoma cells (obtained from Case Western Reserve University and cultured at the Arts lab, University of Western Ontario). The B13 antibody used is a broad neutralizing antibody that binds linear peptide epitope at CD4 binding site of gp120. The cell culture supernatant of the B13 hybridoma was collected and directly applied to the membrane. In the fluorescence system, the fluorophore-conjugated goat anti-mouse monoclonal secondary antibody (Invitrogen) was used in 1:5000 dilution. The membrane was imaged using the Odyssey CLx imaging system (LI-COR Biosciences) at wavelength of 700 nm. In the chemiluminescence system, The HRP-conjugated goat anti-mouse monoclonal secondary

antibody (Invitrogen) was used in a 1: 5000 dilution. The HRP substrate (Immobilon Classico™ from Millipore Sigma) was directly applied to the membrane before imaging on digital scanner (C-DiGit Blot Scanner, LI-COR Biosciences) or on X-ray film (Carestream medical and dental imaging systems) in the dark room.

2.4 Visualization of gp120 using Coomassie

Gels were stained by Coomassie R350 blue solution (0.05% Coomassie blue R350 dye m/v, 25% ethanol, and 10% acetic acid in milliQ water) with gentle agitation overnight. Gels were then de-stained until the protein bands can be distinguished from the clear background.

2.5 Sample preparation for mass spectrometry

Following Coomassie stain, the gel was first washed in ultra-pure water (Milli-Q) twice for 10 minutes each with gentle rocking. The bands of interest were excised and cut into 1 mm^3 cubes. The Coomassie stain was removed by first incubating gel pieces for 15 minutes with 1 volume (volume just enough to cover gel cubes) of ultra-pure water, then another 15 minutes with 1 volume of acetonitrile (Sigma-Aldrich). The water-acetonitrile mixture was removed, and then the gel was incubated in 1 volume of acetonitrile, which was then removed when the gel pieces became shrunk and opaque. Gel was then rehydrated in 1 volume 0.1 M ammonium bicarbonate, pH 8.0 (Sigma-Aldrich) for 5 minutes. One volume of acetonitrile was added to the previous 0.1M ammonium bicarbonate and the gel was incubated in the ammonium bicarbonate-acetonitrile mixture for 15 minutes. The ammonium bicarbonate-acetonitrile mixture was removed. The gel pieces were dried using vacuum centrifuge.

The proteins in gel pieces were reduced by incubation with 1 volume of 10 mM DTT in 0.1 M ammonium bicarbonate at pH 8.0 for 45 minutes at 56 ^{o}C water bath. Proteins were then alkylated by incubation with 55mM iodoacetamide in 0.1M ammonium bicarbonate at pH 8.0 for 30 minutes at room temperature in the dark. The gel pieces were then washed with ultra-pure water and acetonitrile and dried using vacuum centrifuge as previously described.

Gel pieces were rehydrated on ice for 45 minutes with 1.5 volume of digestion buffer (50 mM ammonium bicarbonate pH 8.0, 1 mM $CaCl_2$, 10 ng/µl trypsin (porcine pancreas SOLu-Trypsin 20 ug/ml, Sigma-Aldrich), 10 ng/µl chymotrypsin low (lyophilized bovine pancreas α-Chymotrypsin, ≥40 units/mg, Sigma-Aldrich)). Excess digestion buffer was discarded and replaced with 1 volume of 50 mM ammonium bicarbonate pH 8.0 with 1 mM $CaCl_2$. The gel pieces were incubated at 37 °C overnight to allow for protein digestion.

Liquid containing digested peptides was transferred to a collection tube, to which the liquids from all following incubations were pooled. Gel pieces were then incubated for 15 minutes with 1 volume of 25 mM ammonium bicarbonate at pH 8.0. Without removing the 25 mM ammonium bicarbonate, 1 volume of acetonitrile was added and the gel was incubated for another 15 minutes. The mixture of 25 mM ammonium bicarbonate and acetonitrile was transferred to the collection tube. Gel pieces were next incubated with a mixture of 1 volume of acetonitrile and 1 volume of 5% formic acid for 15 minutes. The acetonitrile-formic acid mixture was transferred to the collection tube. Gel pieces were incubated for the second time with a mixture of 1 volume of acetonitrile and 1 volume of 5% formic acid for 15 minutes. The acetonitrile-formic acid mixture was transferred to the collection tube. Pooled liquid in the collection tube was dried using vacuum centrifuge.

2.6 Mass spectrometry (MS)

Dried samples were submitted to the MS facility (Orbitrap Fusion Lumos) at SPARC BioCentre, SickKids Hospital, Toronto. The samples were resuspended in 12 µl of 0.1% TFA at pH 2.1, from which 6 µl was kept as backup and 6 µl was loaded for liquid chromatography – mass spectrometry (LC-MS). The MS mode was electron-transfer/higher-energy collision dissociation (ETHcD), which is a combination of electron-transfer dissociation (ETD) and higher-energy collision dissociation (HCD) in which HCD triggers ETD upon detection of specific sugar oxonium ions. In this study, the trigger ions were 204.0867 (HexNAc), 138.0545 (HexNAc fragment), 366.1396 (HexNAc-Hex). Detailed mass spectrometry setting information is found in appendix A.

2.7 MS data analysis

One raw dataset was generated per protein band. The raw MS data were converted to mzXML format using a program, MSConvert (downloadable at: http://proteowizard.sourceforge.net/tools.shtml). The mzXML files were analyzed using GlycoPAT (downloadable at: https://sourceforge.net/projects/glycopat/), which was run in MATLAB (downloadable at https://www.mathworks.com/products/matlab.html). Protein fasta sequence, list of potential glycoforms (variable_modifications_custom.txt), batch.m file, and other program files including scoreAll.mako, gather.txt, and fixed_modifications.txt (no content for this study) were fed to GlycoPAT to generate a theoretical library of peptide and glycopeptide ions (digestedpep.txt). The experimental m/z data in the mzXML files were then searched against the theoretical library. An output file (tandems.csv) was generated for each dataset containing glycopeptide assignments in text format for the matched MS spectra. Detailed protocol for the above steps is found in appendix B. Further analysis based on the csv outputs was performed in Microsoft Excel in order to summarize the glycopeptide information in table format and calculate two groups of values: N-Site Distrubution (NSD) and Relative Abundance (RA). NSD for each N-site was calculated by dividing the total number of glycans found at a specific N-site by the total number of glycans found in the entire dataset. RA was calculated by dividing the total number of glycans of a specific glycoform by the total number of glycans found in the entire dataset. This part of the analysis was performed using a semi-automated spreadsheet custom-designed for this project. Detailed protocol containing steps and functions is found in appendix C. Fig. 2 provides a general schematic for this MS data analysis workflow.

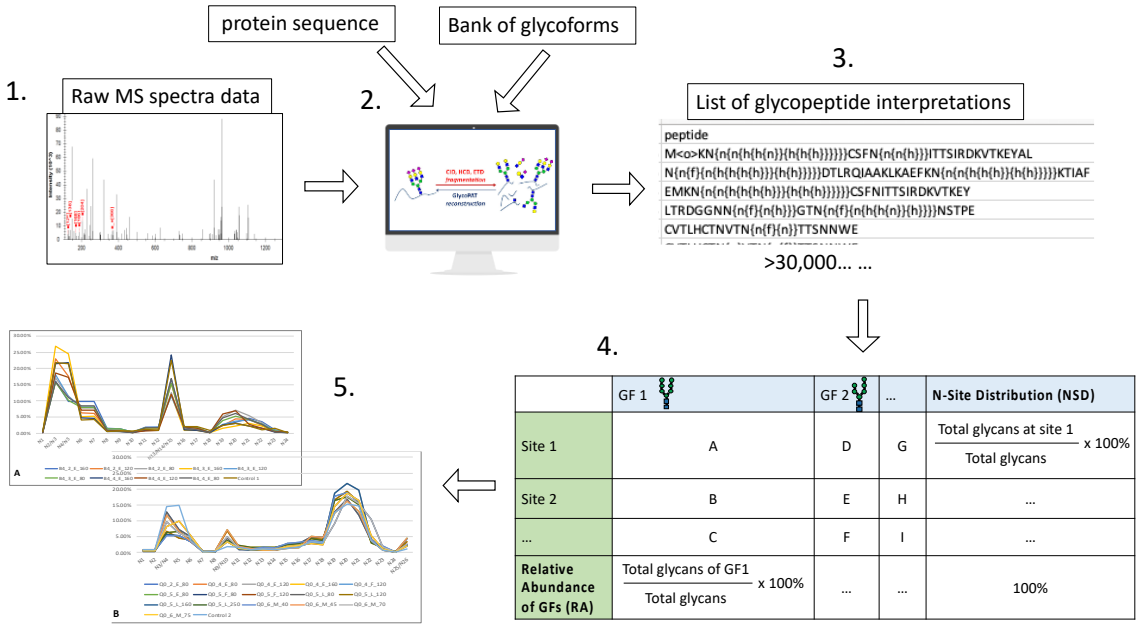


Figure 2: Overall workflow of MS data analysis.

(1). Raw MS datasets were received, each containing thousands of spectra. (2). Using GlycoPAT software, theoretical ion libraries for both B4 and Q0 gp120 sequences were generated using the protein sequence and bank of glycoforms, with inputs of enzymes used for protein digestion. The experimental datasets were then compared with their respective theoretical libraries to look for matched ions. (3). Matched ions were recorded in an csv output file containing a list of the ions' text interpretations, comprising both peptide and glycan information. (4). Each output file was organized and tabulated to calculate the NSD and RA values. (5). The NSD and RA values of all the datasets were visualized in line graphs and analyzed for statistic significance.

2.8 Statistical analysis

ANOVA and Dunnett's post tests with Bonferroni correction were performed based on the NSD and RA values of all the MS datasets of B4 and Q0, regardless of the bands' purification fraction and molecular weight. Unpaired T tests with Welch correction were performed based on the NSD and RA values of a subset of bands (80kDa, elution fraction) that contained three biological replicates for each of B4 and Q0. The statistical

tests were performed using the Prism software (version 9.1.0), downloadable at: https://www.graphpad.com/scientific-software/prism/.

3 Results

3.1 Gp120 purification and selection of protein bands for MS

Three biological replicates of the T/F HIV strain (B4-2, B4-3, and B4-4), and two of the chronic strain (Q0-2 and Q0-3), were first cultured in CD4+CCR5+ U87 cells by Adam Meadows in the Arts lab. The U87 cell line is derived from human glioblastoma cells, which express CD4 but not CCR5 or CXCR4, and therefore not a natural host of HIV. However, the CCR5-transfected U87 cells were shown to outperform the T-cell-derived cell lines expressing CCR5, such as SupT1, MOLT-4 and Jurkat, in supporting HIV replication[118]. We chose the CD4+CCR5+ U87 cell line to produce large amounts of HIV viruses for gp120 glycoprotein purification[118].

Gp120 was purified from the five batches of viruses using a GNA lectin column, as described in section 2.1. Each sample was resuspended in a final volume of 75 µl of 50 mM ammonium bicarbonate solution after dialysis and lyophilization. The samples were visualized by Coomassie stain (Fig.3a) and Western blot (Fig.3b) using 5 µl of each sample.

In the Coomassie gel, the three biological replicates of B4 exhibited an identical pattern of four bands at 160, 120, 80, and 60 kDa. In contrast to B4, the two biological replicates of Q0 showed very different banding pattern. Q0-2 showed a dominant band at 80 kDa and a faint band at 60 kDa, whereas the Q0-3 showed a very strong band at the 55-60 kDa range and many other bands at a variety of molecular weights from 30-250 kDa. As the band pattern and strength of Q0-3 appeared distinct from the rest of the batches, it was excluded from MS processing. Out of the total 120 kDa of gp120, the glycans make up half of this mass and the naked peptide alone is around 60 kDa. In line with other published observations, the 60-kDa band likely represents the non-glycosylated gp120 protein core, or the gag p55 protein[58]. Therefore, this band was excluded from the MS analysis. Therefore, all the 60-kDa bands were excluded as they likely carried few glycans.

In addition to the bands at 120 kDa, bands at 160 and 80 kDa were also selected for MS. The 160-kDa bands matched the mass of gp160, the precursor that forms gp120 and gp41 upon cleavage. The 80-kDa bands were believed to be a truncated fragment of gp120 and consistently appeared in both B4 and Q0, allowing for direct comparison of glycan contents between the two strains. In summary, a total of 9 bands from B4 (160, 120, and 80 kDa of three B4 batches) and 1 band from Q0-2 (80 kDa) were selected to be processed for MS (highlighted in red boxes in Fig. 3a).

Western blot was performed to confirm the identity of the bands in the Coomassie gel (Fig. 3b). In this Western blot using the b13 antibody, many more bands were revealed in comparison to the Coomassie gel due to higher detection sensitivity. This blot was probed with HRP-conjugated secondary antibody, incubated with HRP substrate, and exposed to X-ray film. The image contained several regions with bleached white appearance in the 37-80 kDa range in lanes of B4-2, B4-4, and Q0-2 due to membrane burn-out caused by heat produced by excessive HRP activity, which was an indication of strong signals. Despite this artifact, gp120 was detected by the b13 antibody in all lanes. The presence of lower molecular weight bands was a sign of gp120 degradation during the purification process. The 80-, 120-, and 160-kDa bands selected for MS from the Coomassie gel represented the near-full, full, and precursor lengths of gp120 and were also observed in the Western blot image.

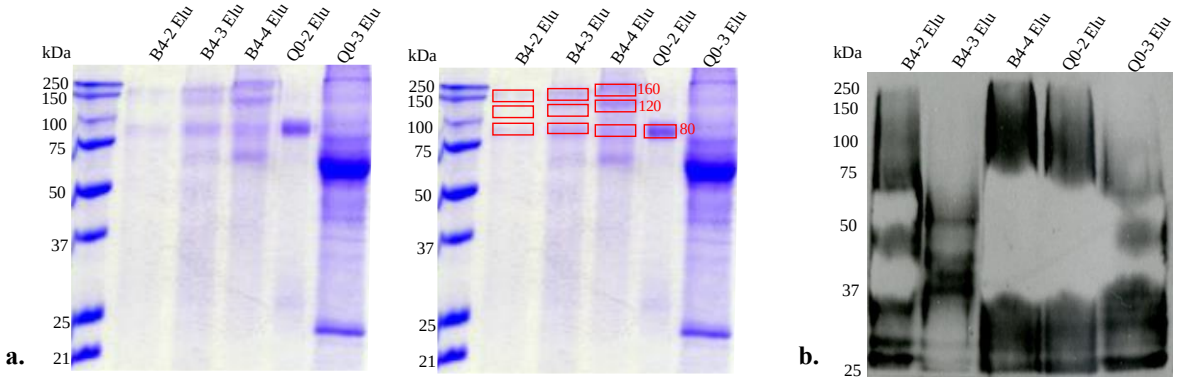


Figure 3: Visualization of the elution fractions of gp120 purification from B4-2, B4-3, B4-4, Q0-2, and Q0-3.

(a). Coomassie stained gel of eluted gp120 samples from the 5 batches. Red boxes highlight the bands cut and processed for MS. (b). Chemiluminescence Western Blot of eluted gp120 samples from the 5 batches using the b13 antibody (binds linear peptide epitope at CD4 binding site of gp120) and HRP-conjugated secondary antibody. The SDS-PAGE protein separation time was longer for the Western blot (b) than the Coomassie stained gel (a), hence the more stretched molecular weight ladder in (b).

To acquire enough data for statistical analysis, two more biological replicates of Q0 virus (Q0-4 and Q0-5) were produced, from which gp120 was purified. Despite the use of the same purification and visualization protocols as the previous batches, the eluted gp120 samples from Q0-4 and Q0-5 produced very faint bands in Coomassie staining (shown in the lanes labeled as “Q0-4 Elu 1” and “Q0-5 Elu”, respectively, in Fig. 4). It was hypothesized that a possible reason for the absence of bands in the elution fractions was that the GNA lectin column used during purification had lost proper binding activity to gp120, such that gp120 was not bound to the column and may be present in the flowthrough. To test this hypothesis, a new column was made, and Q0-4 flowthrough was loaded to re-purify gp120, which produced a new flowthrough and a new elution (shown in the lanes labeled as “Q0-4 FT 2” and “Q0-4 Elu 2”, respectively, in Fig. 4). The flowthrough of Q5 obtained from the first column (shown in the lane labeled as “Q0-5 FT” in Fig. 4) was not re-purified using the new column. Q0-4 Elu 2 showed fainter bands than Q0-4 Elu 1, and both were fainter than Q0-4 FT, which indicated that repurification using new GNA lectin column made no improvements on gp120 concentration in the elution. Therefore, the failure for gp120 to bind the GNA column was not due to the column quality but suggests that the glycosylation pattern of Q0 gp120 might not allow this binding. Western blot (Fig. 4b) was also performed using the same samples as the Coomassie stain (Fig. 4a) for Q0-4 FT, Q0-4 Elu 2, Q0-4 Elu 1, Q0-5 FT, and Q0-5 Elu. The elution fraction from the earlier Q0-3 purification was used as a positive control for the Western blot, and an unrelated 20-kDa bacterial protein was used in two dilutions as negative control.

The membrane was stained with Ponceau S red after transfer and before blocking (Fig. 4b). Like the Coomassie blue stain, the Ponceau S red stain is non-specific and stains all proteins. This Ponceau image showed similar banding patterns as the Coomassie stain (Fig. 4a), except for the 20-kDa bacterial protein controls, which were likely run off the

Coomassie gel. Chemiluminescence-based Western blot (Fig. 4b) was performed using b13 primary antibody (binds linear peptide epitope at CD4 binding site of gp120) and HRP-conjugated secondary antibody. Since bands were very faint in the target molecular weight (120 and 160 kDa), the membrane was exposed to X-ray film overnight. In sharp contrast with the non-specific Coomassie and Ponceau stains, gp120-specific b13 antibody detected higher signal intensities and number of bands in the elution fractions than in the flowthrough. This result was surprising, but the presence of bands in the Q0-3 positive control and the absence of bands in the bacterial protein negative controls were able to provide solid support for the blot. We reasoned that the differential banding patterns in Western blot and Coomassie of the same samples were likely due to the heterogeneity of the gp120 glycans, which affects gp120 interactions with both the GNA lectin column and with the b13 antibody. GNA lectin binds carbohydrates with a nonreducing terminal D-mannose residue. It was possible that only a portion of gp120 had glycans with this particular residue, and the rest were not captured by the GNA lectin column and remained in the flowthrough. Furthermore, b13 antibody detects gp120 by an epitope at the CD4 binding site, which is flanked by glycosylation sites. Depending on their shape and abundance, the glycans could potentially alter or even mask the epitope from b13 detection. The Ponceau S red stain in Fig. 4b showed bands at 60-65 kDa for both elution and flowthrough fractions of Q0-4 and Q0-5. Despite the extremely faint bands, the Coomassie stain in Fig. 4a also revealed a similar pattern. The Western blot in Fig. 4b showed bands in the 60-65 kDa range in the lanes for the elution fractions of Q0-3, Q0-4, and Q0-5. Although Western blot using the b13 antibody did not report bands in this range in the lanes for the flowthrough fractions of Q0-4 and Q0-5, the 60-65 kDa bands of the flowthroughs shown by Ponceau S red and Coomassie were probably gp120 as well, which likely carried different glycoforms from the gp120 in the elutions.

Therefore, it was logical to perform MS on the bands from both flowthrough and elution fractions in order to capture the full spectrum of gp120 glycosylation. However, MS requires that the Coomassie stained bands must be clearly visible to the naked eye, indicating they contain enough quantity of protein. Both flowthrough and elution fractions of Q0-4 and Q0-5 had protein concentrations too low to form clear bands by Coomassie stain with 5 µl of sample loading (Fig. 4a). The Coomassie stain was repeated

for the elution and flowthrough fractions of Q0-4 and Q0-5, and 30 µl of each sample were loaded into the SDS-PAGE wells (maximum loading volume ~20 µl per well) by first drying the samples by vacuum centrifuge and resuspending in a smaller volume of 1x protein loading buffer. With increased sample loading, faint yet visible bands appeared in this Coomassie stain (Fig. 4c).

Although more bands appeared in the new Coomassie stain in Fig. 4c compared to the earlier Coomassie stain in Fig. 4a, the dominant bands at 60 to 65 kDa were again observed. Since the molecular size of the unglycosylated gp120 is around 60 kDa, this band could potentially be the protein core alone and was not chosen for MS because our research interest is the glycosylation signatures. Q0-4 Elu 2 contained no bands, which was consistent with the earlier Coomassie stain (Fig. 3a). Bands at different molecular weights appeared in the flowthrough and elution fractions within the same biological replicate, which aligned with the previous observation that gp120 exhibited a heterogeneity in glycosylation that divided the total protein population into different purification fractions by interacting differently with the GNA lectin column. In summary, this Coomassie gel produced a total of seven bands cut and processed for MS, including one band from Q0-4 FT (120 kDa), three bands from Q0-4 Elu 1 (160, 120, and 80 kDa), two bands from Q0-5 FT (120 and 80 kDa), and one band from Q0-5 Elu (80 kDa).

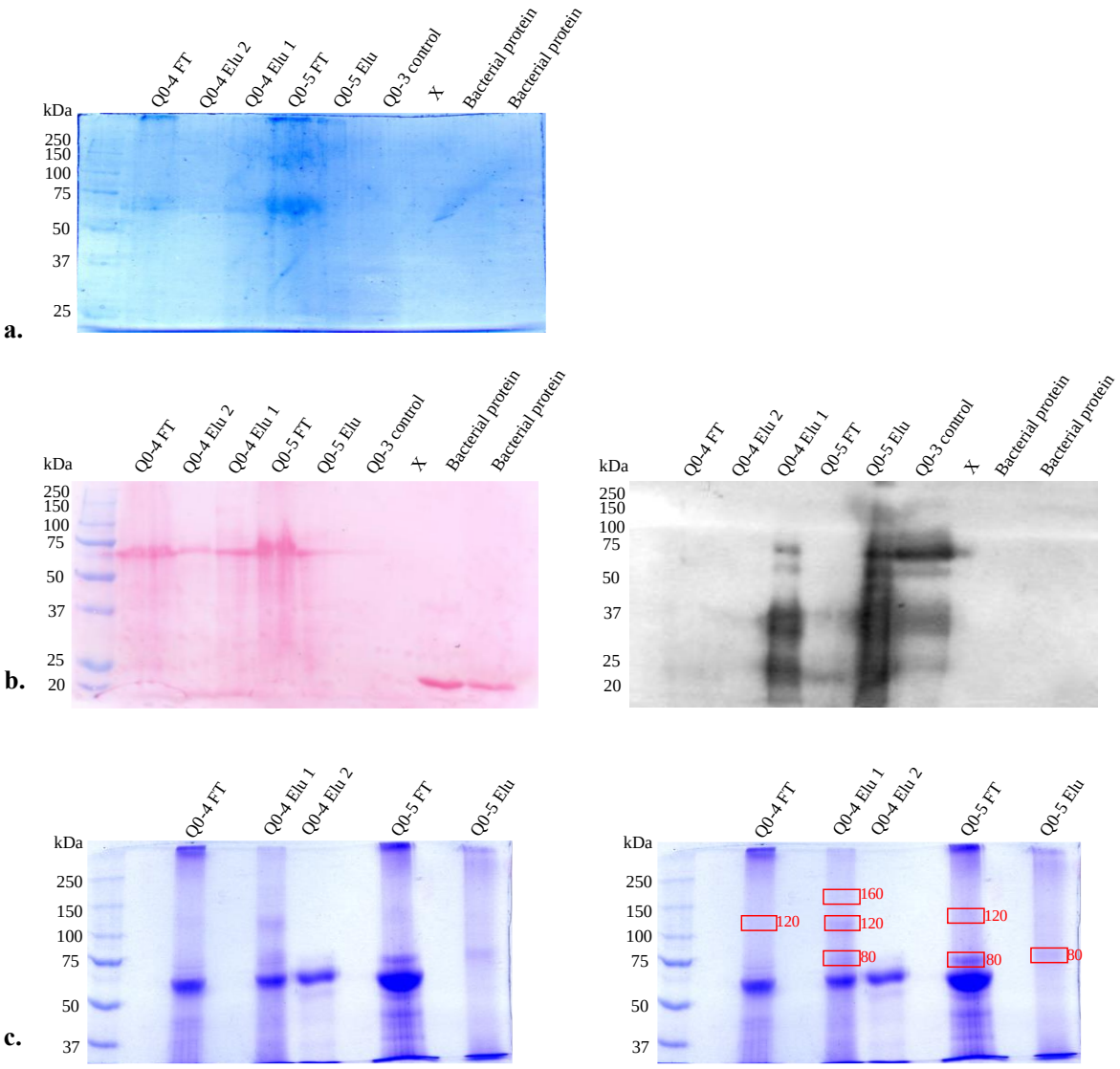


Figure 4: Visualization of the Q0-4 and Q0-5 purification fractions

(a). Coomassie stained gel of Q0-4 and Q0-5 flowthrough and elution fractions with very faint bands. Sample volume = 5 µl. (b). Non-specific Ponceau S red stain of membrane prior to blocking (left) and Chemiluminescence Western blot (right) of Q0-4 and Q0-5 fractions using b13 primary antibody and HRP-conjugated secondary antibody. Sample volume = 5 µl. (c). Coomassie stain of Q0-4 and Q5 fractions with increased sample volume (30 µl). Red boxes highlight the bands cut and processed for MS.

To further investigate the gp120 glycosylation heterogeneity and to acquire more protein sample, another new batch of Q0 (Q0-6) was cultured using the same protocol but 2L of culture supernatant was harvested, which was double the volume of the previous biological replicates. Q0-6 was purified using the same GNA lectin column protocol.

Since the lyophilization step can cause sample loss, the 5 ml of elution fraction was not concentrated by lyophilization after dialysis.

In this Q0-6 purification, knowing the elution and flowthrough fractions represent different subpopulations of gp120, a small volume of viral lysate was kept from passing through the lectin column. The lysate contained the total gp120 population and was used as a standard control to which the elution and flowthrough fractions were compared. In addition, buffer used for washing the column (between flowthrough and elution) was also collected. The lysate, flowthrough, wash, and elution fractions were visualized by Coomassie stain (Fig. 5a) and Western blot (Fig. 5c). Since the samples were not concentrated by lyophilization for the Q0-6 batch, 100 µl of samples were first dried in a vacuum centrifuge and resuspended in 16 µl of 1x protein loading buffer before loading. In the lane named "Q0-6 wash 3v", 300 µl of the wash fraction was used because trace amount of gp120 was expected in this fraction and the 3x volume may help with band visualization. In the Coomassie stain (Fig. 5a), all the fractions, except for the elution, produced a similar banding pattern in various intensities, in which the dominant bands lie between 60-65 kDa. The elution fraction (Q0-6 Elu) showed no clear bands. This pattern was again consistent with the prior purifications of Q0-3 (Fig. 3a), Q0-4, and Q0-5 (Fig. 4a). The lane of Q0-6 lysate served as a total protein control and contained the highest amount of protein, as expected. Q0-6 FT contained the second highest amount of protein after lysate. Q0-6 wash produced much fainter yet clearly visible bands, and the band intensity of the Q0-6 wash 3v lane was as strong as Q0-6 FT, indicating a substantial amount of protein was loosely bound to the GNA lectin column and was lost in the washing steps of the purification process. This Coomassie stained gel suggested that the GNA lectin column fulfilled little function of purification because all the fractions produced the same banding pattern and the band intensity decreases in a stepwise manner from the original lysate to flowthrough, followed by wash, and finally to elution. Therefore, we performed MS on the bands from the lysate fraction. Since the first Coomassie stain showed a smudged appearance in the lane for lysate, it was repeated with ¼ the sample loading (25 µl) to acquire better band quality (Fig. 5b). In this repeated Coomassie stain, the bands appeared much more distinguishable than the original

Coomassie stain (Fig. 5a). Four bands were cut and processed for MS, including 250, 160, 120, and 80 kDa.

Chemiluminescence Western blot was also performed using the same sample volumes and layout as the Coomassie stain in Fig. 5c. B13 primary antibody was used again as previous purifications. Like in the Coomassie stain, Q0-6 lysate produced the strongest bands in Western blot, although the reactive bands are at higher molecular weight than the dominant bands detected by Coomassie staining. Two dominant bands were observed in Q0-6 lysate at the target position of 120 kDa and a high molecular weight above 250 kDa. The high molecular weight band was very likely the intact trimeric form of gp120, undisturbed by the detergent used for viral lysis. The Q0-6 Elu lane contained two major bands at 120 kDa and 60 kDa, both of which were much fainter in intensity compared with the bands in lysate. The absence of the higher than 250 kDa band and the presence of a low 60 kDa in the elution fraction suggested the dissociation of the trimer and the breakdown of monomeric gp120 in the process of purification. Unlike Coomassie, Q0-6 FT and the Q0-6 wash had bands with low signal intensity while Q0-6 Elu had clearly visible bands in Western blot. This differential banding patterns of the Coomassie stain and Western blot were also observed in the previous Q0-4 and Q0-5 purifications.

Both Coomassie stain and Western blot strongly indicated that the purification procedure using the GNA lectin column was not only unnecessary since the glycosylation of Q0 gp120 appeared heterogenous and most of glycosylated Q0 gp120 bound poorly to the GNA resin, but also harmful for the preservation of gp120 quality and quantity. Furthermore, the GNA lectin column also divided the heterogenous gp120 into subpopulations in the flowthrough and elution fractions, leading to a biased selection of bands for MS if only the elution fraction was used to excise protein bands. Besides lectin columns, alternative protein purification systems have been used for gp120 by other researchers, such as immunopurification with a monoclonal antibody and strep-tag affinity chromatography[119-121]. Many studies only use anti-trimer antibodies (e.g., PGT151 and 5F3) in immunopurification of Env[118,119]. In this study, both trimeric and monomeric gp120 are of interest, so the use of monomer-binding antibodies (e.g., 2G12) in addition to the trimer-binding antibodies would be beneficial[28].

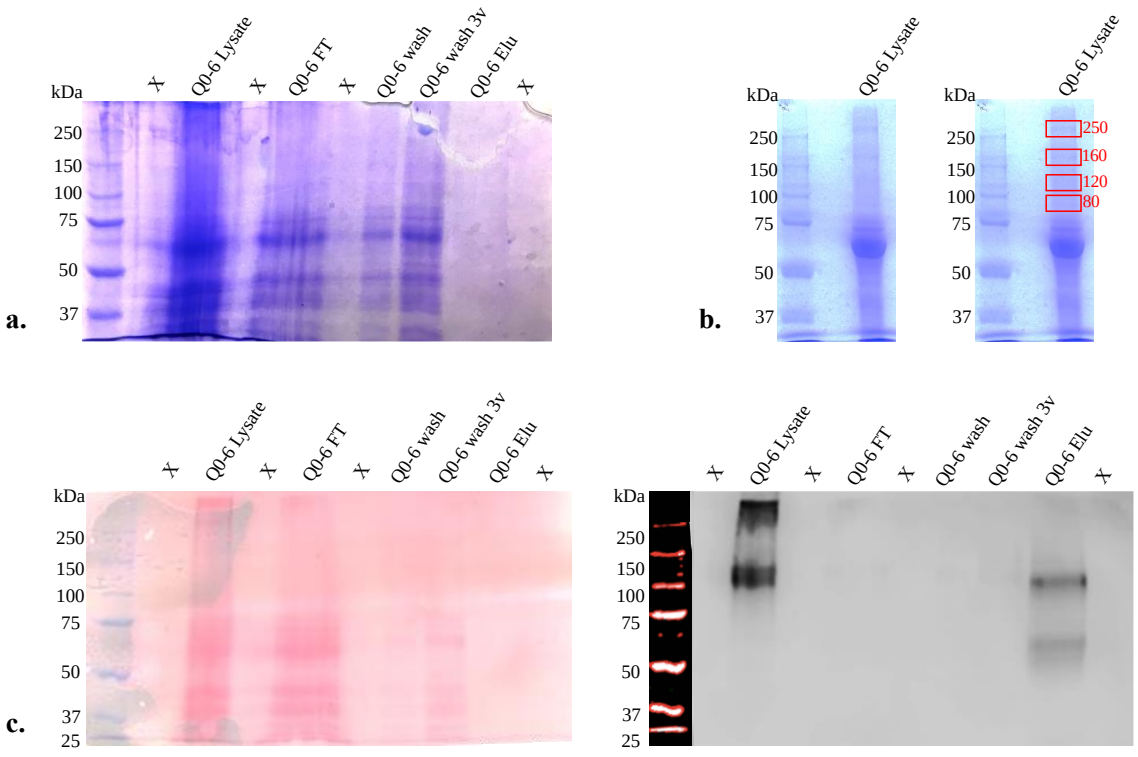


Figure 5: Visualization of Q0-6 purification fractions.

(a). Coomassie stained gel of Q0-6 lysate, flowthrough, wash, and elution fractions. Sample volume = 100 µl. (b). Coomassie stained gel for Q0-6 lysate alone. Sample volume = 25 µl. Red boxes highlight the bands cut and processed for MS. (c). Non-specific Ponceau S red stain of membrane prior to blocking (left) and Chemiluminescence Western blot (right) of Q0-6 fractions using b13 primary antibody and HRP-conjugated secondary antibody. Sample volume = 100 µl. Molecular weight lane of Western blot was imaged using fluorescence scanner while the main image used chemiluminescence scanner to form a composite image.

Unfortunately, the flowthrough and wash fractions of the initial B4 batches were not preserved at the time of purification to be analyzed for any potential gp120 heterogeneity. Therefore, a new batch of B4, B4-5, was produced. Gp120 was purified using the original GNA column method to recreate the flowthrough, wash, and elution fractions, while half the lysate was kept. An additional acid wash fraction, which was the step after elution to remove any residual bound protein from the GNA column by transiently denaturing the GNA lectin, was preserved in this purification. The acid wash fraction was distinct from wash, which occurred between flowthrough and elution. The non-lyophilized fractions were visualized using Coomassie stain (Fig. 6a) and Western blot (Fig. 6b).

In this Coomassie stain, equal volume of each fraction was loaded with the exception of B4-5 wash 3v, which contained 3x the sample volume of the other lanes because trace amount of gp120 was expected in this fraction and the 3x volume may help with band formation. The lanes for B4-5 lysate, B4-5 FT, and B4-5 wash 3v showed clearly visible bands, while the lanes for B4-5 Elu, acid wash, and B4-5 wash 1v showed very faint bands. This pattern was similar to the Coomassie stain of the purification fractions of Q0-6 (Fig. 5a). In the lanes that contained bands, a dominant band at around 50-55 kDa was observed across the lanes, which was lower than the 60-65 kDa dominant band seen in Q0-6 (Fig. 5a). A quadruplet of bands at 250, 160, 120, and 70 kDa was also detected in the same lanes, of which the bands at 250, 160, and 120 kDa likely represented the trimeric form, un-cleaved precursor, and full-length form of gp120, respectively. The lane for B4-5 flowthrough was nearly as stained as the lysate, which was the total protein control, indicating that almost no protein was bound to the GNA lectin column. This further confirmed the conclusion drawn from the Q0-6 purification, which was that the GNA lectin was unnecessary since a majority of gp120 was poorly bound to the column.

In Fig. 6b, the Western blot using b13 antibody revealed gp120 in all fractions of B4-5. This was very different from the Western blot of Q0-6 purification fractions (Fig. 5c), where only the lysate and the elution produced bands. While the dominant bands in the Ponceau stain in Fig. 6b were at around 50-55 kDa, the dominant bands in the respective Western blot were much higher at 120 kDa in all lanes. In addition, the lysate and the flowthrough fractions were very similar in band intensities in both the Ponceau stain (Fig. 6b) and Coomassie stain (Fig. 6a), but the Western blot in Fig. 6b showed much higher signal intensity in the lysate than the flowthrough. B4-5 elution produced no bands in the Coomassie stain but strong bands at 120, 80, and 65 kDa, as well as some faint bands at 55, 40, and 30 kDa. While the flowthrough and wash 3v showed much more band intensities than the elution in Ponceau stain, the bands in the elution fraction appeared stronger than both flowthrough and wash 3v in the Western blot. This was consistent with the Q0-6 fractions as well (Fig. 5), confirming the heterogeneity of gp120 glycosylation in both B4 and Q0.

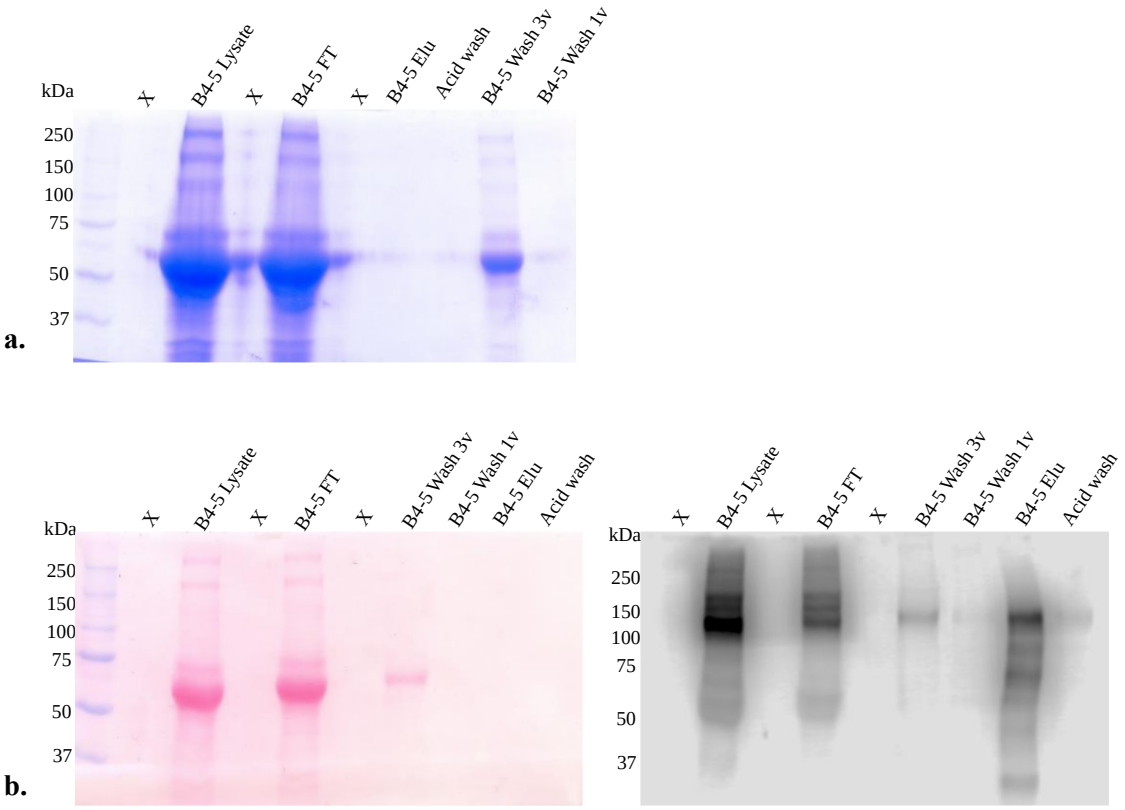


Figure 6: Visualization of B4-5 purification fractions.

(a). Coomassie stained gel of B4-5 lysate, flowthrough, elution, wash, and acid wash fractions. Sample volume = 50 µl, except lane "B4-5 Wash 3v" (150 µl). (b). Non-specific Ponceau S red stain of membrane prior to blocking (left) and Chemiluminescence Western blot (right) of B4-5 fractions using b13 primary antibody and HRP-conjugated secondary antibody. Sample volume = 50 µl, except lane "B4-5 Wash 3v" (150 µl).

3.2 Gp120 in ultracentrifugation supernatant

So far, all the fractions in the original purification procedure used for both B4 and Q0 were examined, including lysate, flowthrough, wash, and elution. However, the collective amount of gp120 in these fractions were low and barely detectable in many cases by Coomassie staining. A portion of gp120 was likely lost in upstream steps prior to purification. After harvesting the viruses in the tissue culture medium and before loading the lysed viral particles onto the GNA lectin column, the viruses were pelleted using ultracentrifugation. The supernatant was discarded in the past because they contained few viruses. However, it was reported in the literature that a substantial proportion of gp120

was observed to shed from the viral envelope into the surrounding milieu[122]. Therefore, the supernatant of ultracentrifugation was kept in order to examine the shed gp120 for the new batches of Q0-6 and B4-5. The supernatant contained 10% FBS, making it rich in proteins such as bovine serum albumin (BSA), which could cause a high level of background noise when studying gp120. Since the molecular weight of BSA is ~66 kDa and gp120 at 120 kDa, Centricon filter tubes with 100-kDa MWCO were used to reduce the levels of such irrelevant proteins. As the flowthrough containing proteins <100 kDa was removed, gp120 was concentrated above the filter with a final concentration factor of 20. Coomassie stain and Western blot were performed for the Centricon-filtered shed gp120 in the supernatant (Fig. 7a and 7b).

This Coomassie stain in Fig. 7a was performed using the centricon-filtered supernatant of Q0-6 showed a dominant band at the 55-65 kDa range, which was likely residual BSA that was too viscous to be removed by the Centricon filters. Other major bands were observed at 160, 120, and 70 kDa. The 160 and 120 kDa bands probably represented the gp160 precursor and gp120, respectively. No Coomassie staining was performed for B4 supernatant as Q0 was of higher priority at the time because MS still required more samples from the Q0 strain.

The fluorescent Western blot in Fig. 7b was performed for the concentrated supernatant of both Q0-6 and B4-5 using the b13 antibody, specific for the CD4 binding site of gp120. B4-5 lysate was used as a positive control since it demonstrated robust reactivity in the previous Western blot shown in Fig. 6b. Q0-6 supernatant and B4-5 supernatant samples were loaded in pairs of different sample volumes of 12, 6, and 3 µl. All the Q0-6 supernatant lanes showed multiple bands in the Ponceau stain with a pattern similar to the Coomassie stain while only one band at 120 kDa produced weak signals in the Western blot. All the B4-5 supernatant lanes showed bands below 75 kDa of the same banding pattern in both Ponceau and Western blot. This indicated gp120 degradation. Indeed, the B4-5 supernatant was acquired earlier than Q0-6 and had suffered an unintended round of thawing and freezing during its storage prior to this Western blot, while the Q0-6 supernatant was freshly prepared. Despite the crooked lanes, the signal intensity of bands of both Q0-6 and B4-5 in this blot were fairly bright given the small amounts of samples

loaded out of the total volume of samples available. There was a total of ~15ml supernatant samples for each of Q0-6 and B4-5 post centricon-concentration, out of which only microliters were used in the Coomassie stain and Western blot. However, this extremely small fraction of supernatant samples yielded clear bands that were comparable, if not stronger, than the bands produced by the samples from prior biological replicates (Q0-4 and Q0-5), which were purified, lyophilized, and resuspended in less than 75 µl of total volume. It was estimated that the total amount of gp120 shed into the supernatant was 100-200 times more than that of the purified gp120 from the viral pellet. This finding was rather surprising; however, it explained the scarcity of purified gp120 from the previous biological replicates.

Since the supernatant represented a substantial amount of shed gp120, it was of interest to analyze this gp120 population using MS. However, before the bands could be processed, the supernatant samples needed to be stored at -80 ^{o}C with 15% glycerol for 3 months during the COVID-19-related closure. After facility reopening, samples were thawed and re-analyzed on Coomassie stain and Western blot to excise bands for MS analyses (Fig. 7c and 7d).

The Coomassie stain and the Western blot images share very similar banding patterns with four dominant bands at 75, 70, 45, and 40 kDa. Compared with the previous Coomassie stain and Western blot of Q0-6 supernatant before the freeze-thaw cycle (Fig. 7a and 7b), the higher molecular weight bands at 120 kDa were no longer present and were replaced with four lower bands instead, suggesting potential protein degradation. This quadruplet of bands for Q0-6 supernatant after sample thawing shown in Fig. 7d closely resemble the bands of B4-5 in Fig. 7b, which was already suffered a cycle of freezing and thawing at the time. Despite the low signal intensity in the Western blot (Fig. 7d), protein bands were excise and processed for MS from the Western blot (Fig. 7d), protein bands were excised and processed for MS from the Coomassie stained gel (Fig. 7c) because MS could detect whether the bands comprised gp120.

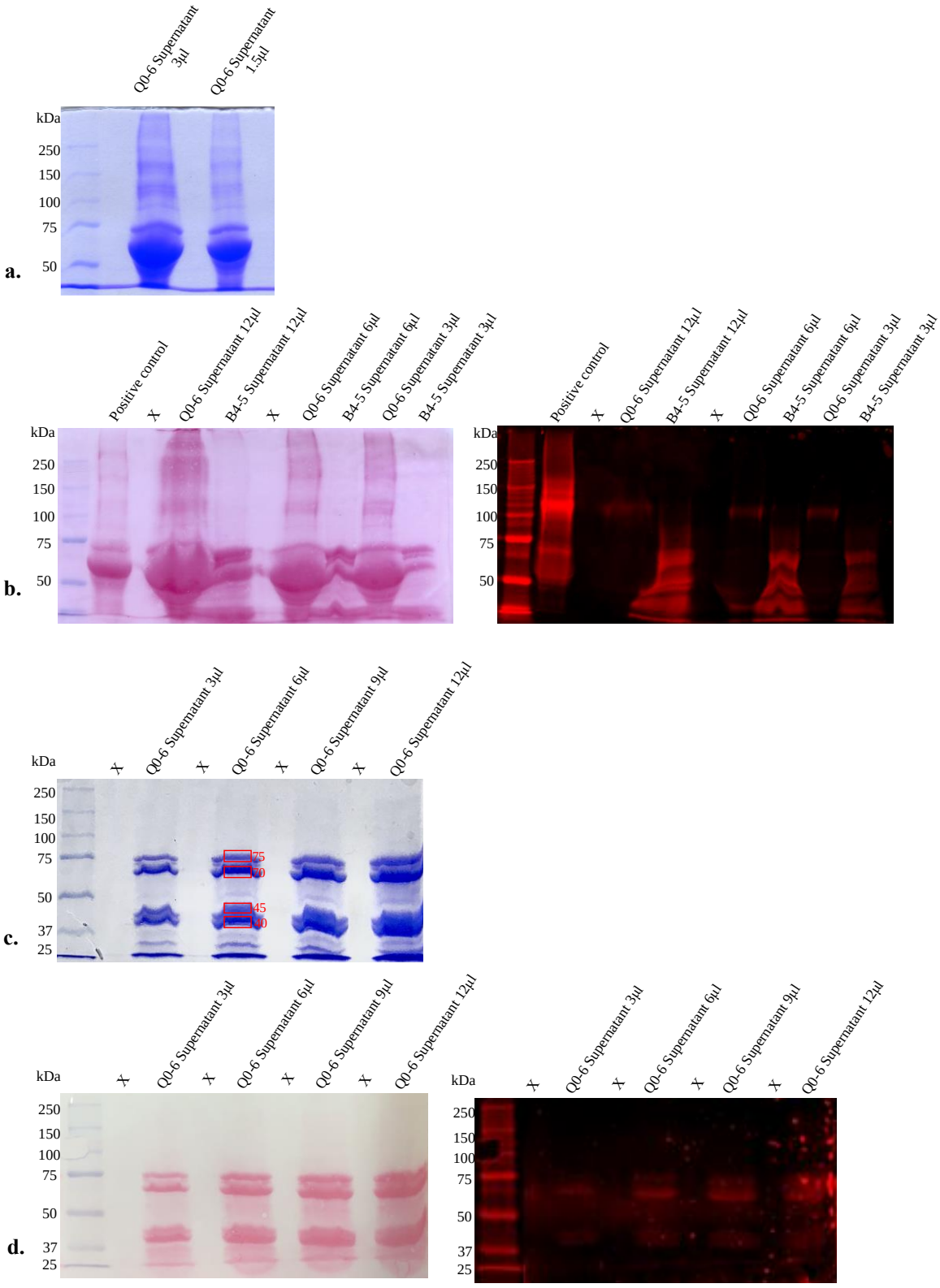

a.
Q0-6 Supernatant 3µl
Q0-6 Supernatant 1.5µl
kDa
250
150
100
75
50
b.
Positive control
Q0-6 Supernatant 12µl
B4-5 Supernatant 12µl
Q0-6 Supernatant 6µl
B4-5 Supernatant 6µl
Q0-6 Supernatant 3µl
B4-5 Supernatant 3µl
kDa
250
150
100
75
50
c.
Q0-6 Supernatant 3µl
Q0-6 Supernatant 6µl
Q0-6 Supernatant 9µl
Q0-6 Supernatant 12µl
kDa
250
150
100
75
50
37
25
d.
Q0-6 Supernatant 3µl
Q0-6 Supernatant 6µl
Q0-6 Supernatant 9µl
Q0-6 Supernatant 12µl
kDa
250
150
100
75
50
37
25

Figure 7: Visualization of shed gp120 in ultracentrifugation supernatant of Q0-6 and B4-5.

(a). Coomassie stained gel of centricon-filtered Q0-6 ultracentrifugation supernatant containing shed gp120. (b). Ponceau S red stain and Western blot (b13 primary antibody and fluorescent second antibody) of centricon-filtered Q0-6 and B4-5 supernatant containing shed gp120. (c). Coomassie stained gel of centricon-filtered Q0-6 ultracentrifugation supernatant after thawing. Protein breakdown (decreased molecular wright of the bands compared to panel a) is observed. Bands boxed in red were cut and processed for MS. (d). Ponceau S red stain and Western blot of Q0-6 supernatant after thawing (b13 primary antibody and fluorescent second antibody).

So far, all possible compartments of gp120 in the Q0-6 viral culture were systematically examined, including supernatant, lysate, flowthrough, wash, and elution. It was determined that the original GNA lectin column purification procedure resulted in the loss of a remarkably large portion of gp120. Choosing bands exclusively from the elution fraction was also proven to be biased. Unfortunately, the other fractions of the initially purified B4-2, B4-3, B4-4, Q0-2, and Q0-3 were no longer available. The elution fractions of the three B4 biological replicates each yielded three bands for MS at 160, 120, and 80 kDa. A newer biological replicate, B4-5, was produced with all fractions preserved and analyzed, confirming that the elution fraction alone could not comprehensively represent the full heterogeneous population of gp120. However, MS has yet to be performed using B4-5 samples because the Q0 strain was a greater priority since the one and only band of Q0 sent for the first batch of MS experiments was the 80-kDa band from Q0-2. In the later biological replicates of Q0-4, Q0-5, and Q0-6, a total of fifteen bands were chosen for MS. Since this batch of MS experiments contained many samples, a naming system for the bands was employed: strain (B4 or Q0), batch number, fraction initial (E for elution, F for flowthrough, L for lysate, and S for supernatant), and molecular weight. The bands were summarized in table 1.

Table 1: Summary of bands sent for MS (bands in blue blocks could be compared directly and statistically because they came from the equivalent fraction and molecular weight)

	B4	**Q0**			
	Elution	Lysate	Flowthrough	Elution	Supernatant
250 kDa		Q0_6_L_250			

160 kDa	B4_2_E_160, B4_3_E_160, B4_4_E_160	Q0_6_L_160		Q0_4_E_160	
120 kDa	B4_2_E_120, B4_3_E_120, B4_4_E_120	Q0_6_L_120	Q0_4_F_120, Q0_5_F_120	Q0_4_E_120,	
80 kDa	B4_2_E_80, B4_3_E_80, B4_4_E_80	Q0_6_L_80	Q0_5_F_80	Q0_2_E_80, Q0_4_E_80, Q0_5_E_80	
40 kDa					Q0_6_S_40
45 kDa					Q0_6_S_45
70 kDa					Q0_6_S_70
75 kDa					Q0_6_S_75

3.3 Three-stage MS data analysis

MS was used to determine the glycosylation signatures of gp120 from B4 and Q0 strains of HIV (schematics shown in Fig. 8). A mode of MS known as electron-transfer/higher-energy collision dissociation (ETHcD, demonstrated in Fig. 8) was used, which is a combination of electron-transfer dissociation (ETD) and higher-energy collision dissociation (HCD) in which HCD triggers ETD upon detection of specific sugar oxonium ions. Specifically, the trigger ions were set as 204.0867 (HexNAc), 138.0545 (HexNAc fragment), 366.1396 (HexNAc-Hex) because these sugars are the most abundant and basic components of gp120 glycans. During MS, digested peptides (parent ions) were each ionized into charged fragments (daughter ions), which were detected by the scanner. Each scan collected the m/z values and intensities of the daughter ions from one parent peptide and produced a spectrum with peaks each representing a daughter ion. In HCD, large glycan structures were poorly preserved and only the base sugars could remain attached to the relatively intact peptide. In ETD, however, the larger glycoforms were more likely to stay intact but the peptide is more fragmented, and the resolution and breadth of detection were less than those of HCD. The combined information of HCD and ETD will be useful to determine both the location of glycan attachment and the specific glycoforms. Fig. 9a and 9b are example spectra of HCD and ETD, respectively.

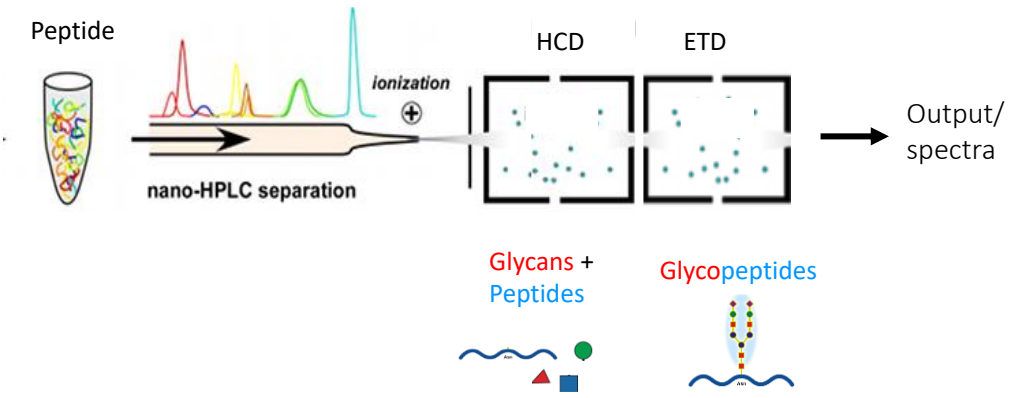


Figure 8: The schematic workflow of ETHcD.

Sample protein is first digested into peptide fragments, which are then separated in high performance liquid chromatography (HPLC) and ionized in the HCD or ETD mode. HCD has higher resolution and breadth of detection than ETD, but only the base sugars could remain attached to the peptide. In ETD, the larger glycoform structures are more likely to stay intact while the peptide is more fragmented. ETHcD, a combination of HCD and ETD, provides both information about both the locations and types of glycans on the peptide chain.

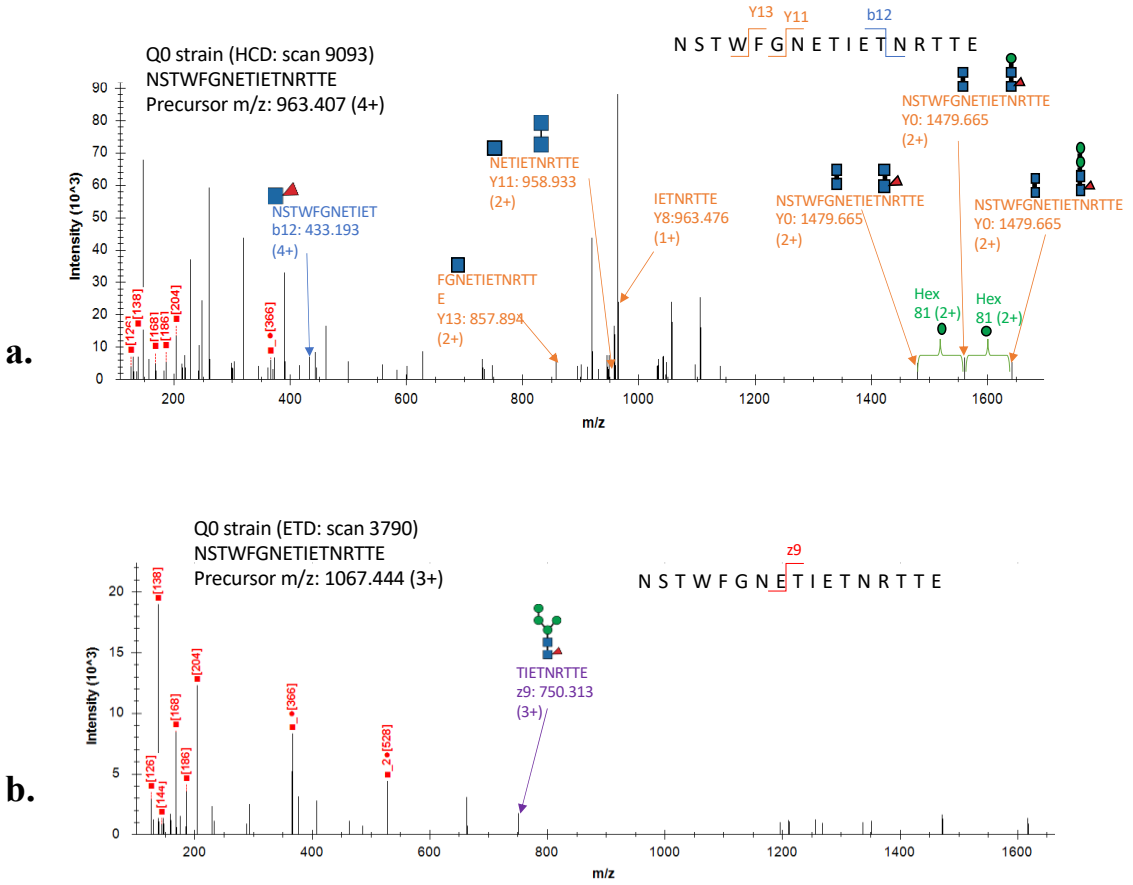


Figure 9: Example spectra from MS raw dataset, manually labelled.
(a). Example HCD spectrum. (b). Example ETD spectrum.

In this manually labelled HCD spectrum, the parent ion produced various B- and Y-types of daughter ions that were color-coded in blue and orange, respectively. The peaks labeled in red on the left side were sugar oxonium ions, which were charged sugar residues cleaved off of the glycans of the parent ion. The gaps between peaks on the right side, which were labelled in green, represented the m/z value of a hexose residue, indicating that the ion peaks were different by one hexose.

When there are multiple N-sites and multiple glycans on a peptide fragment, HCD sometimes could not differentiate which glycan is attached to which N-site. The complementary ETD spectrum may then help locate the glycan to its N-site. For example, the manually labelled ETD spectrum in Fig. 9b showed the same peptide ion in the HCD spectrum (Fig. 9b) with a large glycoform, which is found on the third N-site of this

peptide. This helped to deduce that the largest glycoform in the HCD scan (Fig. 9a) should be labelled on the third N-site. ETD yields C- and Z-types of daughter ions. Fig. 9b shows a Z-type glycopeptide ion (Z9 carrying 3 positive charges, labelled in purple). The ions labelled in red left to the Z9 ion were sugar oxonium ions.

To understand the meaning of the spectra, the m/z values of each daughter ion were matched to a massive theoretical library of m/z values, which was calculated based on the given protein fasta sequence, pool of post-translational modifications (PTMs, details below), and enzymes cleavage sites for protein digestion. The matched daughter ions were then used to deduce the composition of their parent ion in that spectrum. Tens of thousands of spectra were generated for each protein band. To process this enormous amount of data, three stages of analysis were employed: initial computerized search, semi-automated data organizing and cleaning, and statistical analysis. The first two steps required *de novo* data analysis pipeline development which constitute the bulk of the intellectual contribution to this thesis. Detailed protocols on how to proceed are provided in appendices B and C.

3.3.1 Initial computerized search

The raw data in the form of spectra were received from the MS facility and converted to the mzXML format, which could be fed to the GlycoPAT software. As outlined in appendix B, GlycoPAT was used to generate the theoretical library of m/z values based on the known gp120 fasta sequence, a pool of PTMs, and the enzymes used (trypsin and chymotrypsin-low). One library was generated for each of B4 and Q0 fasta sequences. The pool of PTMs contained 29 glycoforms as well as carbamidomethylation and oxidation, potential addition of functional groups during protein band processing. After generating the theoretical library, the experimental m/z values in each spectrum were searched against the library using the GlycoPAT software, leading to one or multiple potential text interpretations of that spectrum. For example, the text interpretation of the spectrum shown in Fig. 9b was: NSTWFGNETIETN{n{f}{n{h{h{h}}{h}}}RTTE. All the spectra were searched against the library, and all the text interpretations were summarized in a csv output file that typically contained 30,000 to 80,000 of lines of

interpretations (methods to analyze these interpretations are provided in appendix C). While some spectra were associated with more than one interpretation, some other spectra contained zero ions that could be matched to the library, and therefore yielded no interpretations at all. The purer the samples were, the higher percentage of spectra could be matched to the theoretical library and interpreted. Since MS is extremely sensitive, protein contamination could generate a high level of noise, especially when the target protein was in scarce amount as in some of the bands submitted, leading to a low percentage of interpreted spectra in the dataset. This percentage in the 9 B4 datasets and 16 Q0 datasets ranged from 10% to 22.5% with a median of 14.3%.

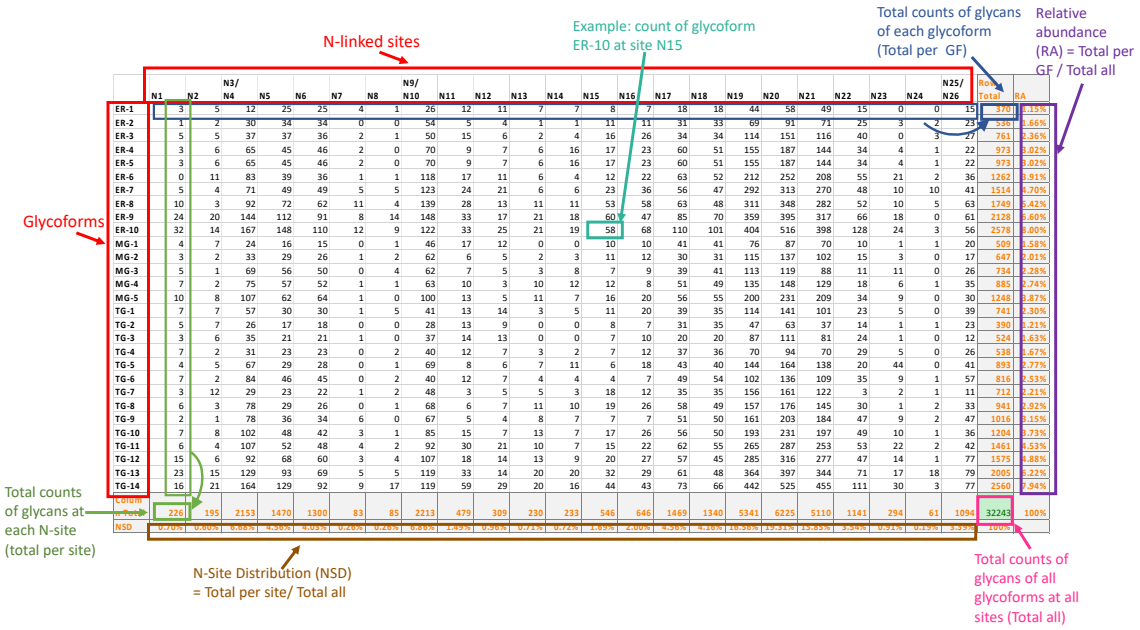
Two control searches (controls 1 and 2) were performed by searching a MS dataset generated under same peptide preparation conditions and the same MS conditions from a bacterial glycoprotein against the B4 and Q0 gp120 libraries, respectively. Since there was no gp120 present at all in the bacterial MS sample, few ions were expected to be matched to the gp120 libraries by chance, leading to very short lists of interpretations in the control outputs. Control 1 showed 6.79% and control 2 showed 6.44% interpreted spectra. Both were lower than the lowest 10.01% among the bona fide experimental datasets and the 14.3% median, but the controls still revealed a substantial level of noise in the current analysis. A possible reason for the higher-than-expected percentage of interpreted spectra in the controls was because this bacterial glycoprotein carried Lewis sugar motifs containing Gal, GlcNAc, Fuc residues, which were common in eukaryotic glycans as well.

3.3.2 Semi-automated data organizing and cleaning

Due to the large sizes of the data, a semi-automated excel-based method was developed specifically for the purpose of organizing and cleaning the interpretations in the csv file, which reduced a task requiring two months of highly repetitive work when performed manually to less than 20 minutes. The development of this method itself was a major milestone in this study, and the stepwise protocol and commends used were described in detail in appendix C. The goal of this stage of analysis was to generate a matrix for each

dataset where the frequencies of each glycoform found at each glycosylation site were summarized (example shown in table 2).

Table 2: Sample Glycoform-Site matrix table from a Q0 dataset.



	N1	N2	N3/ N4	N5	N6	N7	N8	N9/ N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23	N24	N25/ N26	Row Total	RA
ER-1	3	5	12	25	25	4	1	26	12	11	7	7	8	7	18	18	44	58	49	15	0	0	15	370	1.15%
ER-2	1	2	30	34	34	0	0	54	5	4	1	1	11	11	31	33	69	91	71	25	3	2	23	536	1.66%
ER-3	5	5	37	37	36	2	1	50	15	6	2	4	16	26	34	34	114	151	116	40	0	3	27	761	2.36%
ER-4	3	6	65	45	46	2	0	70	9	7	6	16	17	23	60	51	155	187	144	34	4	1	22	973	3.02%
ER-5	3	6	65	45	46	2	0	70	9	7	6	16	17	23	60	51	155	187	144	34	4	1	22	973	3.02%
ER-6	0	11	83	39	36	1	1	118	17	11	6	4	12	22	63	52	212	252	208	55	21	2	36	1262	3.91%
ER-7	5	4	71	49	49	5	5	123	24	21	6	6	23	36	56	47	292	313	270	48	10	10	41	1514	4.70%
ER-8	10	3	92	72	62	11	4	139	28	13	11	11	53	58	63	48	311	348	282	52	10	5	63	1749	5.42%
ER-9	24	20	144	112	91	8	14	148	33	17	21	18	60	47	85	70	359	395	317	66	18	0	61	2128	6.60%
ER-10	32	14	167	148	110	12	9	122	33	25	21	19	58	68	110	101	404	516	398	128	24	3	56	2578	8.00%
MG-1	4	7	24	16	15	0	1	46	17	12	0	0	10	10	41	41	76	87	70	10	1	1	20	509	1.58%
MG-2	3	2	33	29	26	1	2	62	6	5	2	3	11	12	30	31	115	137	102	15	3	0	17	647	2.01%
MG-3	5	1	69	56	50	0	4	62	7	5	3	8	7	9	39	41	113	119	88	11	11	0	26	734	2.28%
MG-4	7	2	75	57	52	1	1	63	10	3	10	12	12	8	51	49	135	148	129	18	6	1	35	885	2.74%
MG-5	10	8	107	62	64	1	0	100	13	5	11	7	16	20	56	55	200	231	209	34	9	0	30	1248	3.87%
TG-1	7	7	57	30	30	1	5	41	13	14	3	5	11	20	39	35	114	141	101	23	5	0	39	741	2.30%
TG-2	5	7	26	17	18	0	0	28	13	9	0	0	8	7	31	35	47	63	37	14	1	1	23	390	1.21%
TG-3	3	6	35	21	21	1	0	37	14	13	0	0	7	10	20	20	87	111	81	24	1	0	12	524	1.63%
TG-4	7	2	31	23	23	0	2	40	12	7	3	2	7	12	37	36	70	94	70	29	5	0	26	538	1.67%
TG-5	4	5	67	29	28	0	1	69	8	6	7	11	6	18	43	40	144	164	138	20	44	0	41	893	2.77%
TG-6	7	2	84	46	45	0	2	40	12	7	4	4	4	7	49	54	102	136	109	35	9	1	57	816	2.53%
TG-7	3	12	29	23	22	1	2	48	3	5	5	3	18	12	35	35	156	161	122	3	2	1	11	712	2.21%
TG-8	6	3	78	29	26	0	1	68	6	7	11	10	19	26	58	49	157	176	145	30	1	2	33	941	2.92%
TG-9	2	1	78	36	34	6	0	67	5	4	8	7	7	7	51	50	161	203	184	47	9	2	47	1016	3.15%
TG-10	7	8	102	48	42	3	1	85	15	7	13	7	17	26	56	50	193	231	197	49	10	1	36	1204	3.73%
TG-11	6	4	107	52	48	4	2	92	30	21	10	7	15	22	62	55	265	287	253	53	22	2	42	1461	4.53%
TG-12	15	6	92	68	60	3	4	107	18	14	13	9	20	27	57	45	285	316	277	47	14	1	77	1575	4.88%
TG-13	23	15	129	93	69	5	5	119	33	14	20	20	32	29	61	48	364	397	344	71	17	18	79	2005	6.22%
TG-14	16	21	164	129	92	9	17	119	59	29	20	16	44	43	73	66	442	525	455	111	30	3	77	2560	7.94%
Column Total	226	195	2153	1470	1300	83	85	2213	479	309	230	233	546	646	1469	1340	5341	6225	5110	1141	294	61	1094	32243	100%
NSD	0.70%	0.60%	6.68%	4.56%	4.03%	0.26%	0.26%	6.86%	1.49%	0.96%	0.71%	0.72%	1.69%	2.00%	4.56%	4.16%	16.56%	19.31%	15.85%	3.54%	0.91%	0.19%	3.39%	100%	

In table 2, the 29 glycoforms were listed in the first column, and the 26 N-linked sites were listed in the first row. Sites N3 and N4 were considered as a site cluster because their close proximity in the protein sequence, which made it difficult to isolate these sites by trypsin and chymotrypsin digestion. Similarly, sites N9/N10 and sites N25/N26 were clustered as well in all Q0 datasets. However, B4 protein sequence was different from Q0 and its N-sites were clustered differently, as shown in table 3 below. Cells in the second last row contained the total number of glycoforms found at the N-site in their respective column. The green cell at the end of this row contained the total number of glycoforms across all N-sites in this dataset. The row below contained the percentages of the glycoforms found at a specific N-site out of the total glycoforms in this dataset, also referred to as **N-site distribution, or NSD.** The second last column contained the total number of a specific glycoform across all N-sites. The last column contained the percentages of a specific glycoform out of all glycoforms, referred to as **relative abundance, or RA**.

All the MS output files, including those of the controls, were organized into tables of this format and standardized into NSD and RA percentages, thereby directly comparable across different biological replicates and strains. Before statistical analysis could be performed, the 2-dimensional matrix tables of all the datasets were collapsed into two 1-dimensional tables each focusing on either NSD (tables 3 for B4 and table 4 for Q0) or RA (tables 5 for both B4 and Q0). Furthermore, these 1-dimensional tables of all the datasets were visualized using line graphs (Fig. 10 and 11).

Table 3: N-site distribution (NSD values) of each site for all B4 datasets including control 1 (bacterial protein data searched against B4 library).

All values were in %. NSD is the percentage of the glycans found at a specific N-site out of the total glycans in the respective dataset. NSD reflects the spatial pattern of glycan distribution along the peptide chain within each independent dataset.

	N1	N2/N3	N4/N5	N6	N7	N8	N9	N10	N11	N12	N13/N14/N15	N16	N17	N18	N19	N20	N21	N22	N23	N24
B4_2_E_160	0.36	16.01	9.77	9.80	9.88	1.58	1.43	0.39	1.94	1.84	16.72	1.39	1.08	0.37	2.60	3.37	4.52	3.59	0.96	0.15
B4_2_E_120	0.27	22.93	17.63	6.26	6.15	1.12	1.29	0.23	1.34	1.35	15.31	1.16	1.07	0.19	2.47	4.40	4.28	2.41	0.48	0.20
B4_2_E_80	0.19	17.05	11.43	7.80	7.87	1.08	0.91	0.28	1.49	1.49	12.50	1.45	1.16	0.16	5.10	7.17	5.49	3.24	0.70	0.30
B4_3_E_160	0.34	26.93	24.47	5.17	5.23	0.73	0.82	0.21	1.36	1.31	11.84	0.92	0.85	0.14	1.50	2.24	3.15	2.29	0.47	0.20
B4_3_E_120	0.63	18.07	11.26	8.09	8.21	1.39	1.37	0.34	1.74	1.87	11.50	2.12	1.91	0.20	2.81	3.78	4.15	2.96	0.90	0.27
B4_3_E_80	0.53	15.67	10.15	7.90	7.98	1.32	1.42	0.40	1.75	1.76	15.34	1.93	1.74	0.27	3.99	5.29	4.17	2.78	0.71	0.37
B4_4_E_160	0.16	21.41	21.83	4.70	4.66	0.53	0.56	0.27	0.86	0.86	24.16	1.04	0.88	0.13	2.51	3.07	2.24	1.53	0.33	0.19
B4_4_E_120	0.25	18.54	17.30	7.08	7.06	0.94	0.68	0.42	0.76	0.82	12.04	1.92	1.56	0.22	5.98	6.94	2.69	1.51	0.53	0.30
B4_4_E_80	0.18	15.83	10.94	8.51	8.47	0.64	0.57	0.26	1.74	1.69	16.93	1.75	1.08	0.25	4.73	6.03	4.17	2.75	0.41	0.24
Control 1	0.59	21.75	21.54	4.13	4.29	0.62	0.67	0.75	1.24	0.84	22.55	2.07	2.04	0.92	2.34	3.13	2.24	1.01	1.52	0.24

Table 4: N-site distribution (NSD values) of each N-site for all Q0 datasets, including control 2 (bacterial protein data searched against Q0 library).

All values were in %. NSD is the percentage of the glycans found at a specific N-site out of the total glycans in the respective dataset. NSD reflects the spatial pattern of glycan distribution along the entire peptide chain within each independent dataset.

	N1	N2	N3/N4	N5	N6	N7	N8	N9/N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23	N24	N25/N26
Q0_2_E_80	0.36	0.42	12.81	7.18	4.89	0.08	0.24	4.61	0.99	0.88	0.84	0.81	2.82	2.87	4.91	4.76	12.71	16.73	11.85	3.58	1.13	0.13	4.40
Q0_4_E_80	0.70	0.60	6.68	4.56	4.03	0.26	0.26	6.86	1.49	0.96	0.71	0.72	1.69	2.00	4.56	4.16	16.56	19.31	15.85	3.54	0.91	0.19	3.39
Q0_4_E_120	0.49	0.47	5.73	5.38	3.44	0.14	0.15	3.53	1.19	0.81	0.85	0.81	1.76	1.97	2.99	2.75	18.65	21.81	19.57	4.19	0.80	0.13	2.40
Q0_4_E_160	0.81	0.84	6.29	6.72	4.03	0.28	0.36	3.82	2.17	1.73	1.75	1.69	2.49	2.90	4.32	3.87	16.42	17.55	15.02	2.94	1.52	0.26	2.22
Q0_4_F_120	0.84	0.71	5.34	5.34	3.48	0.33	0.27	3.45	1.67	1.19	1.70	1.45	2.90	3.23	3.84	3.30	17.78	18.94	16.09	3.92	2.05	0.20	1.97
Q0_5_E_80	0.83	0.84	11.80	6.59	3.56	0.24	0.31	7.30	1.89	1.39	0.73	0.70	1.40	1.59	5.09	4.80	12.36	16.25	12.99	3.77	0.82	0.22	4.54
Q0_5_F_80	0.75	0.77	9.78	6.12	4.80	0.34	0.33	4.97	1.53	0.94	1.13	1.19	1.73	1.97	3.32	2.97	9.18	17.71	15.46	10.41	1.67	0.24	2.70
Q0_5_F_120	0.57	0.49	8.11	9.86	5.28	0.25	0.27	3.46	1.33	0.96	0.91	0.96	1.89	2.27	2.80	2.38	14.55	18.74	16.36	5.26	1.08	0.15	2.06
Q0_5_L_80	1.28	0.82	5.34	6.50	6.01	0.50	0.55	2.43	2.19	1.60	1.33	1.44	2.36	3.04	4.38	3.39	14.92	17.88	15.93	4.71	1.13	0.13	4.40
Q0_5_L_120	1.35	1.07	8.09	9.50	6.76	0.44	0.74	3.93	2.39	1.86	2.08	2.07	2.50	2.69	5.03	4.47	12.53	14.45	10.93	3.01	0.91	0.19	3.39
Q0_5_L_160	1.56	1.13	11.91	12.89	5.88	0.71	1.05	2.58	3.06	2.78	2.44	2.46	2.53	3.97	8.42	6.44	6.93	8.48	7.53	3.22	0.80	0.13	2.40
Q0_5_L_250	1.63	1.23	9.23	9.56	6.96	0.53	0.88	2.96	3.00	2.67	2.49	2.71	3.03	3.67	6.58	5.15	9.36	11.64	9.08	3.42	1.52	0.26	2.22
Q0_6_M_40	1.20	1.16	6.31	8.62	7.53	0.61	0.75	3.96	1.56	1.74	2.87	2.35	2.85	3.25	3.44	2.89	11.01	12.77	8.89	7.58	2.05	0.20	1.97
Q0_6_M_45	0.89	0.97	10.85	6.55	4.13	0.49	0.45	4.01	1.06	1.45	2.60	1.82	2.00	2.52	3.36	2.70	6.58	14.80	11.73	11.57	0.82	0.22	4.54
Q0_6_M_70	0.68	0.83	12.76	7.02	4.55	0.38	0.44	4.80	1.16	1.14	1.71	1.45	1.79	2.21	3.54	3.04	9.48	15.90	11.97	9.03	1.67	0.24	2.70
Q0_6_M_75	0.74	0.81	11.69	6.49	4.11	0.32	0.33	3.70	1.49	0.99	1.36	1.12	1.62	2.12	3.94	3.33	13.45	17.59	13.99	6.81	1.08	0.15	2.06
Control 2	0.84	1.01	14.48	14.94	3.85	0.24	0.42	1.93	1.60	0.89	1.03	1.02	1.37	1.61	3.39	3.05	12.61	15.23	14.74	3.79	0.77	0.14	1.08

Table 5: Relative abundance values (RA) of each glycoform for all datasets including two controls.

All values were in %. RA is the percentages of a specific glycoform out of all glycoforms.

	ER-1	ER-2	ER-3	ER-4	ER-5	ER-6	ER-7	ER-8	ER-9	ER-10	MG-1	MG-2	MG-3	MG-4	MG-5	TG-1	TG-2	TG-3	TG-4	TG-5	TG-6	TG-7	TG-8	TG-9	TG-10	TG-11	TG-12	TG-13	TG-14
B4_2_E_160	1.63	2.12	3.18	3.46	3.46	4.06	4.32	5.90	7.43	8.16	1.64	2.11	2.84	2.95	3.25	2.03	1.71	1.60	2.42	2.22	2.75	1.94	2.41	2.65	3.29	3.79	4.36	5.95	6.36
B4_2_E_120	2.80	3.22	3.49	3.53	3.53	3.84	4.19	4.72	5.87	6.07	2.98	3.46	3.12	3.57	3.97	2.37	2.82	2.02	3.19	2.27	2.69	1.94	2.75	2.46	3.14	3.66	3.42	4.25	4.66
B4_2_E_80	2.45	2.69	2.97	3.49	3.49	4.01	4.06	4.82	5.16	6.01	2.00	2.12	2.61	2.90	3.26	2.71	2.35	2.25	2.83	2.42	3.30	1.82	3.21	2.93	4.01	4.41	4.86	5.41	5.41
B4_3_E_160	2.19	2.28	2.60	2.83	2.83	3.50	4.74	5.58	7.55	7.78	1.57	1.77	1.73	2.89	3.32	1.84	2.10	1.57	2.26	1.86	2.75	1.59	2.48	2.89	3.21	4.47	5.32	7.07	7.44
B4_3_E_120	2.25	2.51	3.53	3.52	3.52	4.12	4.78	5.77	7.37	8.51	1.79	2.01	2.02	2.76	3.57	1.71	1.98	1.35	2.30	1.95	2.66	1.54	2.69	2.47	3.07	3.72	4.42	5.98	6.12
B4_3_E_80	2.11	2.55	3.31	3.43	3.43	4.19	4.96	5.69	6.73	7.96	1.77	2.05	2.06	2.59	3.29	1.60	1.85	1.26	2.38	2.02	2.09	1.57	2.82	2.52	3.19	4.42	5.02	6.20	6.93
B4_4_E_160	3.07	3.31	3.21	3.55	3.55	4.00	4.60	5.27	4.82	6.57	3.11	2.84	2.24	2.39	3.65	1.89	2.41	2.50	2.63	2.62	1.88	2.82	2.69	2.81	3.21	4.41	4.73	4.30	4.91
B4_4_E_120	2.03	2.66	2.84	3.09	3.09	4.14	4.60	5.41	5.49	7.66	2.21	2.37	1.89	2.40	3.35	1.85	2.04	1.93	2.34	2.29	2.42	2.75	2.54	3.38	3.64	4.55	5.38	5.19	6.49
B4_4_E_80	1.54	2.41	2.18	2.98	2.98	3.53	4.75	5.64	6.05	7.52	2.01	2.03	2.75	3.40	3.46	2.08	1.79	1.44	1.91	1.82	2.74	1.71	1.93	3.05	3.28	4.99	5.85	6.24	7.92
Control 1	2.82	2.94	3.14	3.53	3.53	3.83	4.47	5.17	6.01	8.60	2.10	2.04	2.25	2.44	2.87	2.16	2.86	2.21	2.78	2.11	2.30	2.46	3.17	2.50	2.82	2.83	4.18	4.57	7.31
Q0_2_E_80	1.65	2.29	2.72	3.05	3.05	3.76	4.83	6.23	7.99	8.94	1.98	1.85	1.94	2.39	3.04	1.80	1.58	1.65	1.81	2.18	2.14	1.76	2.29	2.61	3.10	3.72	5.19	6.67	7.77
Q0_4_E_80	1.15	1.66	2.36	3.02	3.02	3.91	4.70	5.42	6.60	8.00	1.58	2.01	2.28	2.74	3.87	2.30	1.21	1.63	1.67	2.77	2.53	2.21	2.92	3.15	3.73	4.53	4.88	6.22	7.94
Q0_4_E_120	1.19	1.65	2.16	3.24	3.24	4.11	4.98	5.62	6.14	7.74	1.71	1.96	1.90	2.16	3.78	1.84	1.64	2.19	2.06	3.08	2.26	1.82	3.31	3.47	4.18	4.97	4.98	5.48	7.13
Q0_4_E_160	1.34	1.71	2.06	2.67	2.67	3.93	4.93	5.84	6.82	8.33	1.83	2.00	2.17	2.50	3.94	1.69	1.41	1.93	1.96	2.52	2.04	2.58	2.46	3.13	3.67	4.77	5.23	6.39	7.47
Q0_4_F_120	1.14	1.43	1.93	2.68	2.68	3.85	4.67	5.86	6.86	8.95	1.52	2.00	1.95	2.06	3.67	1.72	1.52	2.02	2.13	2.57	2.12	2.57	2.56	3.17	3.85	4.79	5.26	6.51	7.93
Q0_5_E_80	1.36	1.62	1.88	2.86	2.86	3.46	4.62	6.76	8.26	9.68	1.23	1.41	2.30	2.21	2.92	1.88	1.29	1.32	1.42	2.44	2.07	1.44	2.78	2.55	2.93	4.25	5.89	7.03	9.28
Q0_5_F_80	1.52	1.99	2.82	3.53	3.53	4.14	4.71	5.71	6.46	8.35	1.55	2.10	1.85	1.90	3.01	1.89	1.61	1.88	2.31	2.51	2.01	2.06	2.96	2.59	3.57	4.85	5.38	6.37	6.86
Q0_5_F_120	1.33	2.04	2.51	3.26	3.26	4.24	4.95	5.34	6.24	6.85	2.12	2.37	1.89	2.11	3.58	1.77	1.76	2.29	2.31	2.77	1.95	1.45	3.11	3.35	4.20	4.91	5.15	6.21	6.71
Q0_5_L_80	2.30	2.59	3.17	3.76	3.76	4.06	4.83	5.54	6.85	8.68	1.87	1.95	1.31	1.31	2.36	1.24	2.12	1.93	2.59	2.31	1.25	1.14	3.03	2.18	3.39	4.44	5.33	6.27	8.43
Q0_5_L_120	1.27	1.66	2.09	2.85	2.85	3.48	4.90	5.98	8.61	9.87	1.16	1.73	1.34	2.14	3.00	1.44	1.04	1.29	1.67	2.11	2.07	1.56	2.37	2.56	3.14	4.66	5.57	8.17	9.43
Q0_5_L_160	1.18	1.20	2.14	3.06	3.06	3.72	5.07	6.87	8.59	11.37	0.89	1.35	1.53	1.79	2.55	1.45	1.20	1.18	1.66	2.10	1.60	1.71	2.60	2.29	3.56	4.18	5.16	6.98	9.95
Q0_5_L_250	1.33	1.54	2.49	3.12	3.12	3.74	4.76	6.46	9.14	10.90	1.20	1.53	1.48	1.93	2.91	1.49	0.99	1.43	1.51	2.11	1.82	1.11	2.38	2.32	3.03	4.21	5.35	7.16	9.46
Q0_6_M_40	1.26	1.91	2.09	3.29	3.29	4.38	5.78	6.07	7.34	10.63	1.36	1.42	1.30	1.76	2.79	0.98	1.60	1.39	1.93	1.87	1.52	1.27	2.42	2.17	3.47	4.92	5.86	6.87	9.06
Q0_6_M_45	1.95	2.41	2.82	3.77	3.77	4.24	5.61	5.43	6.22	9.17	1.43	1.69	1.36	1.74	2.49	1.53	1.94	1.70	2.39	2.31	1.59	1.17	3.28	2.42	4.02	4.47	5.30	5.98	7.83
Q0_6_M_70	2.03	2.63	3.18	3.77	3.77	4.19	5.43	5.69	5.96	8.38	1.77	2.25	1.84	2.17	3.06	1.58	1.86	1.71	2.49	2.16	1.86	1.82	2.87	2.40	3.39	4.32	4.77	5.41	7.22
Q0_6_M_75	1.78	2.41	2.73	3.17	3.17	3.82	4.88	5.96	6.15	8.78	1.93	2.40	1.80	1.94	2.85	1.46	1.87	1.85	2.44	2.53	1.84	2.18	2.98	2.65	3.43	4.89	4.87	5.67	7.56
Control 2	2.79	3.51	3.59	3.64	3.64	3.39	3.47	3.53	5.73	7.02	2.58	2.66	2.32	2.44	3.32	2.92	2.41	2.58	2.50	2.99	3.06	1.84	3.01	3.25	3.11	3.36	3.86	5.27	6.21

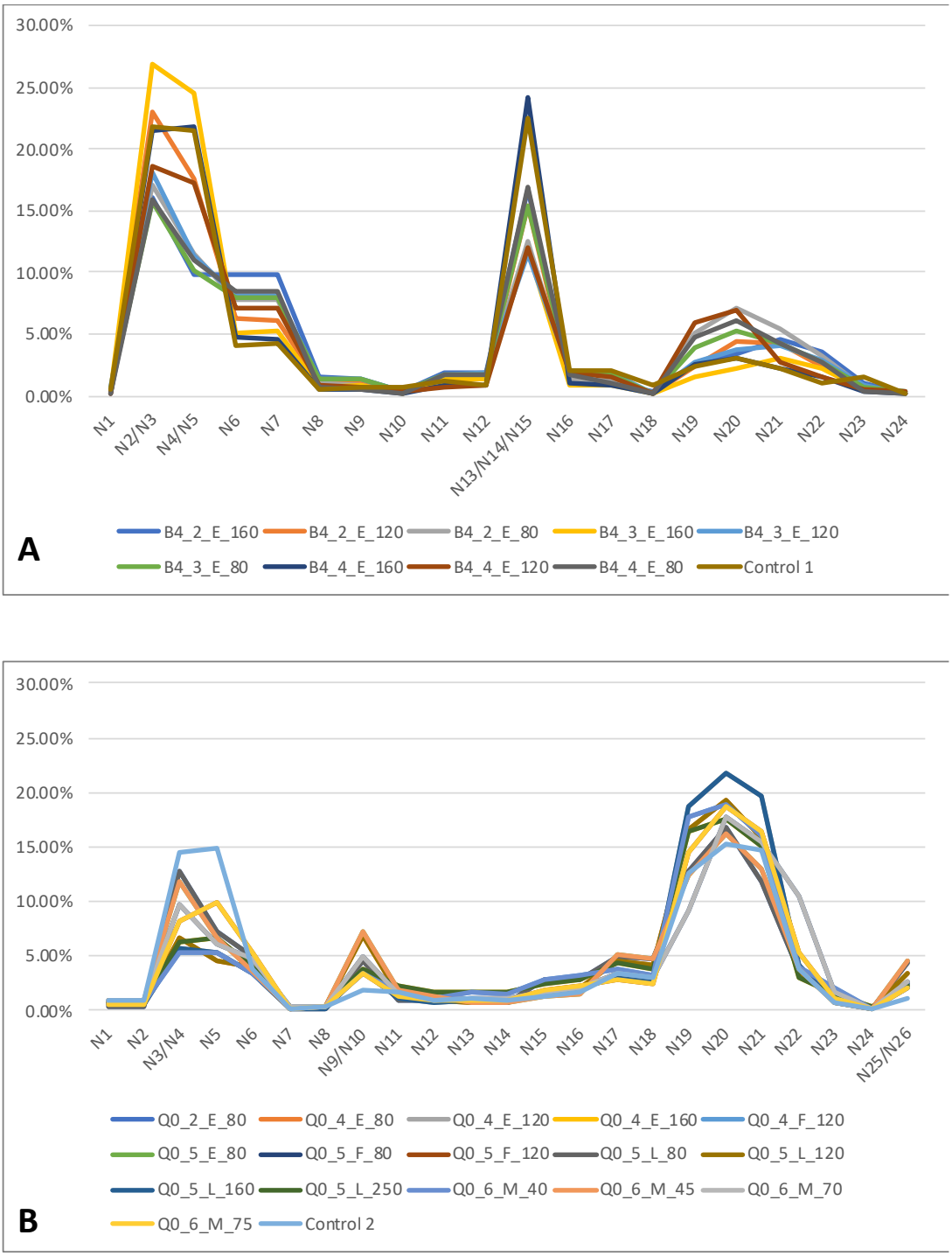
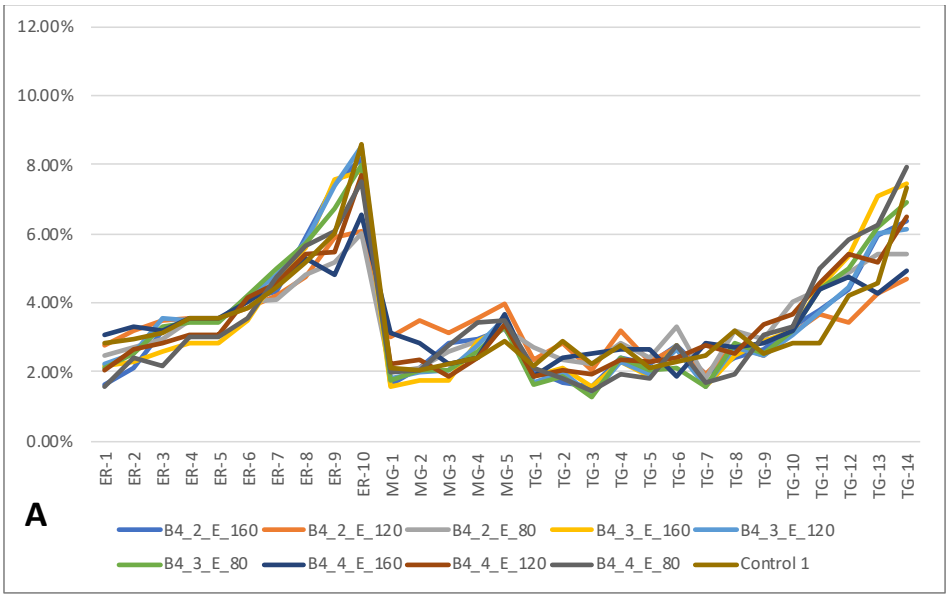


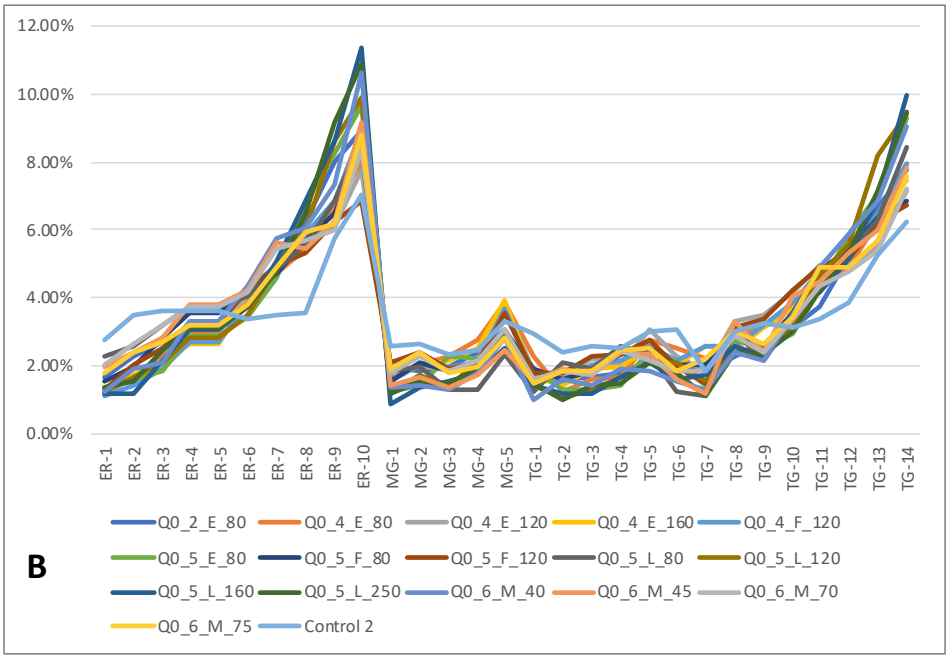
Figure 10: Line graphs of the N-site distribution (NSD) values of all B4 and Q0 bands.

(A). NSD values of each N-site along the gp120 sequence across all B4 bands and control 1. (B). NSD values of each N-site along the gp120 sequence across all Q0 bands and control 2. All values were in %. NSD is the percentage of the glycans found at a specific

N-site out of the total glycans in the respective dataset. NSD reflects the spatial pattern of glycan distribution along the entire peptide chain within each independent dataset.

Panels A and B of Fig. 10 showed the NSD of glycans across all B4 datasets and Q0 datasets, respectively. B4 and Q0 showed apparent differences in their glycan density distributions across the protein length. B4 had the most glycan density from sites N3 to N6 near the N-terminus and from sites N13 to N15, a cluster of N-sites in the middle portion of the peptide chain. Large proportions of Q0 glycans were also found from sites N3 to N6 but the largest peak of glycans locate from sites N19 to N22. Control 1 largely followed the general trend of NSD of the B4 datasets, while control 2 deviated from the general trend of NSD of the Q0 datasets at N3-N5 and N19 to N21.





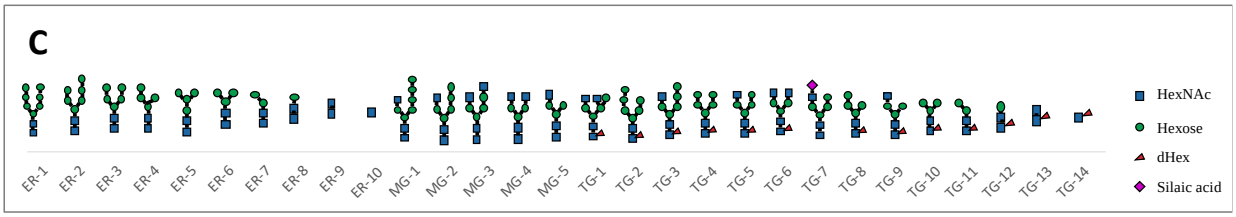


Figure 11: Line graphs of the relative abundance (RA) values of the glycoforms in all B4 and Q0 bands.

(A). RA values of the glycoforms across all B4 Bands and control 1. (B). RA values of the glycoforms across all Q0 Bands and control 2. C. Visual representation of all the glycoforms included in this study. RA is the percentages of a specific glycoform out of all glycoforms.

Fig. 11c showed the structures of the 29 glycoforms included in this study. Human cells could produce many more possible glycoforms. However, including more glycoforms can create tremendous computational burden. As a proof-of-principle study, we chose to include these glycoforms because they are among the most basic and frequently detected in MS. In future studies, more glycoforms should be included to achieve higher resolution of glycopeptide screening to identify potential epitopes for vaccine design.

Glycoforms ER1-10 were placed in the endoplasmic reticulum (ER) block since they only contain hexose residues in their branches, also known as high-mannose glycoforms. Nascent forms of glycans from the ER progressively mature as they were transported to the *medial* Golgi (MG) and *trans* Golgi (TG), where HexNAc and dHex residues were added to the branches, forming hybrid and complex branches. Specifically, the most abundant HexNAc in this study is typically N-acetylglucosamine, and the most abundant dHex is fucose.

Panels A and B of Fig.11 showed the global RA of glycoforms, which described the percentage of each of the 29 glycoforms out of the total glycans in the dataset globally, independent of N-sites. B4 and Q0 exhibited very similar trends where the smaller-sized glycoforms, such as ER-9 and TG-14, were detected in larger proportions than the larger glycoforms. This similarity was explained by the fact that both viruses were produced by the same cell line, and glycosylation was largely cell-type dependent. It could also be explained by the significant fractionation of the glycoforms during the MS analysis performed, which was needed to map the glycosylation site, but resulted in enrichment in small glycoforms. Despite the overall similarity, B4 appeared to have a more even selection of the glycoforms than Q0 and contained higher proportions of large-sized glycans in the ER block. This suggested that the B4 gp120 was less processed than the Q0 counterpart during post-translational modification in the Golgi, and this difference in gp120 glycosylation may be linked to B4’s enhanced infectivity. Given that only one

strain was examined for each of T/F and chronic HIV in this study, further research on many different acute vs T/F strains would be required to confirm this speculation.

3.3.3 Statistical analysis

ANOVA and Dunnett's post tests have been performed for the NSD values on the 9 B4 datasets and the 16 Q0 datasets. The numerical site names (e.g., N1, N2, etc.) were based on the sites' positions from the N-terminus in B4 or Q0 protein sequence. Having the same numerical name did not necessarily indicate homology between the B4 and Q0 sites. Prior to statistical analysis, the homologous pairs of N-sites of B4 and Q0 were aligned, shown in table 6.

Table 6: Alignment of B4 and Q0 N-sites by protein sequence homology.

B4 site	Q0 site	Homologous site pair (HSP)	P values
N1	N2	HSP 1	>0.9999
N2/N3	N3/N4	HSP 2	0.9993
N6	N5	HSP 3	0.1771
N7	N6	HSP 4	<0.0001
N8	N7	HSP 5	0.9982
N9	N8	HSP 6	0.9985
N10	N9/N10	HSP 7	<0.0001
N11	N11	HSP 8	0.9997
N12	N12	HSP 9	0.9989
N13/N14/N15	N13	HSP 10	<0.0001
N16	N15	HSP 11	0.9996

N17	N16	HSP 12	0.9981
N18	N17	HSP 13	<0.0001
N19	N19	HSP 14	<0.0001
N20	N20	HSP 15	<0.0001
N21	N21	HSP 16	<0.0001
N22	N23	HSP 17	0.1719
N23	N24	HSP 18	0.9983
N24	N25/N26	HSP 19	<0.0001

The statistical analysis showed that there was a significant difference in glycan density distribution ($P<0.0026$) between B4 and Q0 at the homologous site pairs (HSPs) highlighted in green. Since 19 HSPs were compared and alpha inflation must be considered, the P value cut-off of 0.0026 was calculated by dividing 0.05 by 19, according to the Bonferroni correction for multiple comparisons. These HSPs indicated the key N-sites where B4 and Q0 exhibited differences in glycosylation signatures, which may be linked to their differential abilities to transmit. This finding brought forth a great future direction for research, where more strains of T/F and chronic HIV viruses may reveal a more definitive correlation or potential causation between N-site glycosylation and transmission.

Since Q0 samples originated from 4 different sources (lysate, flowthrough, elution, and supernatant), while B4 samples all emerged from the elution fraction only, it was logical to specifically examine the B4 and Q0 datasets of the same fraction and molecular weight. The bands at 80 kDa from elution fractions of B4 (B4_2_E_80, B4_3_E_80, and B4_4_E_80) and Q0 (Q0_2_E_80, Q0_4_E_80, and Q0_5_E_80) are the only two groups that can be directly compared and with three biological replicates (n=3) to acquire meaningful statistics (band information summarized in table 1). To clarify, in this case, n is equal to 3 in both B4 and Q0 since the comparison is exclusively between the two

particular strains and each strain contained gp120 samples purified from three independent batches of viruses. However, in a larger scale study to compare T/F and chronic viruses in general, all bands from the same strain would only constitute as n=1 in statistical tests where multiple strains would be included for both the T/F and chronic groups.

As mentioned in section 3.1, prior to MS, these 80-kDa bands were believed to be a truncated form of gp120, which could be a result of protein degradation during the purification process. This is later confirmed since gp120-specific ions were found in the MS datasets. However, the entire gp120 sequence was covered by glycopeptide ions in the MS result, so we were unable to distinguish the exact location of this 80-kDa fragment on the gp120 sequence. Unpaired T tests with Welch correction were performed for the NSD values (the percentage of the glycoforms found at a specific N-site out of the total glycoforms) of the two groups of bands. Three homologous site pairs (HSPs 13, 15, and 17) were discovered to have significant differences in their NSD values ($P<0.0026$). In the statistical test using all the datasets (results shown in table 6), In the statistical test using all the datasets (results shown in table 6), more HSPs were found to show significant differences in their NSD values, including HSPs 4, 7, 10, 13, 14, 15, 16, and 19. However, HSP 17 showed significance in the sub-group of bands (80kDa bands from elution fraction) but not in the calculation using all bands, suggesting that the bands of different molecular weight potentially contained different fragments of gp120 with different glycosylation signatures. This further emphasizes the need to perform comparisons between carefully matched samples in terms of sample origin and processing step.

After NSD values were evaluated, the global RA values (the percentages of a specific glycoform out of all glycoforms in the respective dataset) of the two groups of 80 kDa elution bands were also examined using unpaired T tests with Welch correction. None of the 29 glycoforms showed significant differences in global RA between the B4 and Q0 groups. Next, site-specific RA values (the percentages of a specific glycoform out of all glycoforms at the respective site) were examined using the same test for HSP 13, 15, and 17 individually. Like the global RA values, the T tests of HSP 13-specific, HSP 15-

specific, and HSP 17-specific RA values showed no glycoforms that are significantly different between the B4 and Q0 groups. The non-significant statistic results from global and site-specific RA values, in combination with the significant results from the NSD values, suggested that the locational distribution of glycans had greater impacts on the B4 and Q0 glycosylation signatures than the relative abundance of specific glycoforms, at least in the 80 kDa bands from the elution fraction. However, since the sample size in this comparison was small, this preliminary finding requires further investigation with more repetitions and in bands of other molecular weights and from different fractions.

4 Discussion and Conclusions

In this study, we set out to develop a workflow to analyze MS data of the gp120 samples purified from a T/F strain (B4) and a chronic strain (Q0) in order to characterize the unique structural features of T/F HIV viruses that may supporting their strong transmission fitness.

Before the MS analysis, gp120 was purified from cultured viruses using GNA lectin columns. We found that there was an innate heterogeneity of gp120 within each batch of virus. This was demonstrated repeatedly in all batches tested by the presence of a large proportion of gp120 that did not bind to the column (even when the binding capacity of the column was not exceeded) and by the differential banding patterns in Coomassie stain and Western blot using b13 antibodies, where the elution fraction showed very faint bands in Coomassie stain but strong bands in Western blot while the flowthrough fraction showed the opposite pattern. Since the b13 antibody detects gp120 by a peptide epitope at the CD4 binding site (CD4bs), which is flanked by glycosylation sites, b13's ability to detect gp120 could be potentially reduced if this epitope is masked or altered by the dense glycan shield surrounding the CD4bs. It was possible that the gp120 in the elution fraction contained less glycans than its counterpart in the flowthrough fraction, such that their epitope was more accessible to the b13 antibody. To investigate this issue, gp120 with various point mutations that disable the N-sites near the CD4bs could be produced and visualized in Western blots using the b13 antibody. If the epitope for b13 can indeed be masked by nearby glycans, removing N-sites should increase b13 signal in Western blots.

Gp120 in the elution fraction was barely detected using Coomassie, despite the strong signals in Western blot using b13 antibody. Coomassie staining is non-specific and less sensitive than Western blot, indicating that this population of gp120 only constituted a small proportion of the total gp120. The majority of gp120 were found in the flowthrough and were likely highly glycosylated at N-linked sites surrounding the b13 epitope, resulting in weak band intensities in the accompanying Western blots. This structural heterogeneity within the gp120 population not only explained the different reactivities to

the b13 antibody, but also the different binding activities to the GNA lectin column in the first place. The GNA lectin only binds carbohydrates with a nonreducing terminal D-mannose residue. It was likely that only a small portion of gp120 had glycans with this particular residue, and the rest were not captured by the GNA lectin column and remained in the flowthrough. Alternatively, the higher level of glycosylation of gp120 recovered in flowthrough may have masked the high-mannose patch region, preventing the gp120 molecules from binding to the column. In summary, we found that the total gp120, whether sourced from the B4 strain or the Q0 strain, constituted heterogenous subpopulations that differ at least in the density of glycans near the b13-specific epitope and near the high-mannose patch and/or in the abundance of glycoforms with nonreducing terminal D-mannose residues. To further study how different glycan structures may contribute to this heterogeneity in gp120's ability to bind b13 and GNA lectin, Western blot experiments using antibodies with glycan epitopes, such as 2G12 and PGT121, may be helpful[28,123].

Furthermore, the heterogeneity of gp120 could be a result of either the cell type (CD4+CCR5+ U87) used to culture the viruses, or certain viral accessary proteins that may interfere the host glycosylation machinery, or a combination of both factors. To investigate whether the observed gp120 heterogeneity is intrinsic to the viruses, other cell types may be used and the resulting viral gp120 examined. Previous studies have produced SOSIP Env trimers and viral Env trimers using cell lines such as the Chinese hamster ovary (CHO), 293F, 293T, and the peripheral blood mononuclear cells (PBMCs)[124-126]. These studies compared the degree of glycan processing of the SOSIP and viral Env trimers, however they did not report the intra-strain heterogeneity of gp120 glycans.

Regardless of the cause of the gp120 heterogeneity, the purification procedure using the GNA lectin column should be avoided in future studies involving total gp120 because it could lead to unnecessary sample loss and introduce bias in the subsequent analyses by separating the total gp120 into elution and flowthrough fractions. Neither of these fractions alone could comprehensively represent the properties of the total gp120, and the whole viral lysate should be used as MS samples. GNA lectin-based gp120 purification

method was initially described by Gustav Gilljam in 1993 and widely used in gp120 studies[127]. However, we could not find any studies that examined the flowthrough for gp120 heterogeneity. The use of gp120 from the elution fraction alone could be a limitation of those studies as their conclusions about gp120 glycosylation were based on a non-representative subpopulation of gp120 molecules. As previously suggested in section 3.1, antibody-based immunopurification and strep-tag affinity chromatography may be used as alternative methods to purify gp120[118-120]. With immunopurification, the use of both trimer- and monomer-binding antibodies for gp120 would be beneficial to increase the yield of protein samples.

We also found that a substantial amount of gp120 was shed from the viral particles into the surrounding tissue culture medium. After harvesting the culture medium containing live viruses, the medium was first centrifuged to remove the cellular debris and then the cell-free supernatant was spun again using ultracentrifugation to pellet the viruses. The viral pellet was used as the raw material for gp120 purification, while the virus-free supernatant from the ultracentrifugation was typically discarded. However, we not only discovered gp120 in the virus-free supernatant, but also estimated that the supernatant contained 100-200 times more gp120 in the supernatant than in the viral pellet based on the band intensities from Coomassie stain and Western blot (Fig. 7). This finding largely challenged the standard practice of studying gp120 from the viral pellet alone. An interesting study has shown that shedding-resistant Env can be made by either creating mutations at the gp160 precursor's protease cleavage site or introducing disulfide bridge between gp120 and gp41 subunits[128]. Therefore, in the future we could employ these Env-stabilizing techniques to reduce shedding and largely increase yield of the purified glycoprotein for MS.

The gp120 was likely released from the virions into the supernatant in two stages: during tissue culture and during ultracentrifugation. Gp120 shedding from HIV particles, as well as other retroviruses, was long observed by many researchers and was one of the mechanisms of viruses to neutralize the antibody response[98]. We hypothesize that the gp120 in the supernatant was likely to be derived from the ultracentrifugation process, in which large g forces (100 000x g) was applied to the sample and could damage some

viral particles, freeing the envelope-bound gp120 into a soluble form. If the large g force can indeed cause viral structural damage, 15% sucrose cushion may be helpful to reduce this negative effect. However, we chose not to use sucrose since it is a sugar, and may therefore interfere with the lectin-based gp120 purification. In future studies, 15% sucrose cushioning may be used with washing steps to increase the gp120 yield from the viral pellet.

It was unclear whether the shed gp120 in the supernatant had any structural differences from the gp120 in the viral pellet, and if so, whether these structural differences can promote or prevent shedding. We processed the protein bands for MS from the Coomassie stained gel using supernatant samples. However, no definitive conclusion can be drawn yet since there was only one set of supernatant bands from one biological replicate of Q0. Further research on this subject should consider processing supernatant samples for MS from more biological replicates and other strains of HIV.

Our main research focus in this study was to use MS to compare the glycosylation signatures of gp120 from B4 and Q0 HIV strains. The MS mode used was ETHcD, a combination of HCD and ETD. In HCD, large glycan structures were poorly preserved and only the base sugars could remain attached to the relatively intact peptide backbone, providing crucial information on glycan attachment location. In ETD, the larger glycoforms were more likely to stay intact but the peptide is more fragmented, providing information on glycoform type. The combined information of HCD and ETD is useful to determine locations and types of glycans in B4 and Q0 strains. MS was performed for 3 biological replicates of gp120 from acute HIV (B4 strain) and 4 biological replicates of chronic HIV (Q0 strain). Each biological replicate involves up to 4 protein bands from various fractions, including elution, flowthrough, lysate, and supernatant, as shown in Table 1. MS raw data are in form of spectra containing peaks that represent daughter ions fragmented from a parent ion. The position of a peak on the horizontal axis depicts the m/z value of the daughter ion, and the height depicts the relative intensity. Analysis of MS data requires a pre-calculated theoretical library of m/z values of all possible daughter ions that can be generated from the given protein sequence, pool of potential post-translational modifications, and enzymes used to digest the full-length glycopeptide

into shorter fragments, which become the parent ions before further MS fragmentation. Each spectrum is analyzed by matching the peaks against this theoretical library and deducing the possible parent ion, translating the spectra from picture to a peptide or glycopeptide assignment, which is a peptide sequence potentially carrying glycans noted in standardized glycobiology nomenclature. When all the spectra in a dataset are processed, it is then possible to determine the presence and abundance of particular glycoforms at particular N-linked sites based on the matched spectra. A typical MS dataset from one protein band consists of around 30,000 spectra. The calculation of the theoretical library and the subsequent matching between the experimental m/z values to the library were both performed using the GlycoPAT software in the Matlab environment. After the matching process is completed for each dataset, an output file containing the peptide and glycopeptide assignments translated from the matched spectra is generated. This computational method was first made available by Gang Liu *et* [129] in 2014 and employed in the pilot study of this project by Dr. Najwa Zebian, a former member of the Creuzenet lab. The output of this step requires further analysis, cleaning and summarizing the text into tables and graphs with biological meaning. Due to the novelty of the study and massive size of high-throughput data, there was originally no established method to effectively achieve this goal. Throughout this thesis project, an efficient excel-based semi-automated workflow was developed (detailed protocol available in appendix B) to summarize the information from an entire dataset into a site-vs.-glycoform matrix table. The essential steps of this workflow involved first sorting the glycopeptide assignments in the order of their appearance in the gp120 sequence, then grouping the assignments with the same N-sites and glycoforms, and finally generating a summary table showing the frequency of each glycoform at each N-linked site. This table made it possible to present biological meaning of each dataset and to run statistical tests using multiple datasets. This workflow was not only useful for gp120 studies but could also be utilized for any other studies involving MS of glycoproteins, such as an ongoing project in our lab about bacterial glycoproteins.

After all the datasets were summarized in tables, statistical tests were run using the N-site distribution (NSD) values and the relative abundance (RA) values of each dataset. The NSD and RA values are the standardized percentages calculated by dividing the number

of glycans at a specific site and the number of glycans of a specific glycoform, respectively, by the total number of glycans found in that dataset. An ANOVA test was first performed using the NSD values of all datasets of B4 and all datasets of Q0, regardless of the molecular weight and fraction of the bands from which the datasets were derived. Since B4 and Q0 have different protein sequences and different positions of N-sites, the test was only able to compare the NSD values of the 20 homologous site pairs (HSPs), which are the N-sites common in both B4 and Q0. This test was run to find the general differences in the spatial distribution of glycans at each N-site on the B4 and Q0 peptide backbone. The ANOVA test showed that B4 and Q0 significantly differ in the NSD values at 8 HSPs (HSPs 4, 7, 10, 13, 14, 15, 16, and 19) among the total of 20 HSPs compared, with 4 of the 8 sites (13, 14, 15, and 16) concentrated in the V4 region of the gp120 sequence, which is part of the outer domain in folded gp120. While this ANOVA test compared the NSD values (indicating glycan distribution by sites) of the B4 and Q0 bands, no statistical test was performed for the RA values (indicating abundance of each glycoform) as they showed nearly identical trends in line graphs. This similarity was expected since the two HIV strains were both cultured in the same cell line. However, the RA values of some high-mannose glycoforms (processed in the endoplasmic reticulum) were slightly higher in B4 than in Q0, while some complex type glycoforms (processed in the *medial*- and *trans*- Golgi) showed slightly higher abundance in Q0 than in B4. Although the differences in glycoform RA values were small between B4 and Q0, this trend provided an interesting future direction to study whether B4 could escape some later steps in glycan maturation in the Golgi, therefore leaving the high-mannose ER glycoforms, which are more nascent, on gp120.

Unlike the first general ANOVA test using all the datasets available, a second statistical test (T test with Welch correction) was performed using only the NSD values from 3 datasets of B4 and 3 datasets of Q0 because these emerged from proteins bands of the same purification fraction and molecular weight (Elution_80kDa), allowing for a direct comparison between the two viral strains. The T test indicated that 3 HSPs (HSPs 13, 15, and 17) had significantly different NSD values between B4 and Q0. HSPs 13 and 15 also showed significance in the first general ANOVA test using all bands, while HSP 17 did not, suggesting that the bands of different molecular weights and fractions potentially

contained different fragments of gp120 with different glycosylation signatures. More samples should be processed for MS such that direct comparisons between carefully matched bands can be made.

Similar to the T test performed for the NSD values of the 3 pairs of matched bands, another T test was performed for the global RA values (each glycoform's percentage out of all glycans of the entire dataset) from the same datasets. No glycoform showed significant difference between B4 and Q0. Next, 3 more T tests were performed using the site-specific RA values (each glycoform's percentage out of all glycoforms at a specific N-site) of HSPs 13, 15, and 17. Like the previous test using the global RA values, the site-specific RA values showed no significant differences between B4 and Q0 either.

In summary, the significant results from the NSD values and the non-significant results from global and site-specific RA values together suggested that the spatial distribution of glycans was a greater factor than the glycoform types to determine the overall glycosylation signatures in B4 and Q0. Since the sample size in this study was small, this preliminary finding requires further investigation with more repetitions and in bands of other molecular weights and from different fractions. However, the HSPs with significantly different glycan contents provided a good starting point to find glycopeptide vaccine targets on gp120 in future studies.

No study is without its limitations. There are 4 main limitations in this study. First, only one strain was studied for each of the T/F and chronic groups. HIV is notorious for its frequent mutations, leading to numerous strains and subtypes. Only one strain per group could not accurately represent the general characteristics of the two groups. This project was conceived as a proof of principle study, and thus now enables completion of broader studies in the future with more HIV strains. Second, the bands processed for MS from the B4 strain only sourced from the elution fraction. As discussed earlier, elution alone could not reflect the full heterogeneous population of gp120. Although this limitation was corrected in the Q0 bands, B4 would still require more samples for MS to match those of Q0. Third, during the MS data analysis, the control using a bacterial glycoprotein revealed a high background noise in the MS data. This was likely due to two reasons: 1,

there was a scarce amount of protein samples in the faint Coomassie stained bands; 2, the control dataset and the bona fide gp120 bands shared common sugar residues and a few structurally unique glycoforms with the same masses, which cannot be distinguished by MS. To overcome this limitation, future studies could use a blank gel piece as control, rather than another glycoprotein. Finally, only 29 basic glycoforms, some of which were the highly fragmented forms, were included in this study, while the human cell machinery could produce hundreds of glycoforms. A larger pool of potential glycoforms would certainly improve the quality of the theoretical library of ions and increase the number of matched MS spectra; however, it can also add tremendous computational burden and require the use of high-performance computing (HPC). Since operating the HPC requires substantial computer science (CS) knowledge, a collaboration project between the microbiology and CS departments would be a great way to improve the breadth of this study.

Overall, this thesis examined the standard gp120 purification protocol and found that, rather than the elution fraction, the lysate fraction is best able to represent the heterogenous gp120 population. Therefore, MS protein bands should be cut from the lysate fraction in future experiments. The analysis workflow tools developed in this thesis will also enable prompt analysis of all future datasets in directly comparable formats to elucidate the glycosylation signature (glycoforms and attachment locations) of gp120 that contributes to the HIV transmission fitness.

References

1. Zaitseva M, Blauvelt A, Lee S, Lapham CK, Kiaus-Kovrun V, Mostowski H, Manischewitz J, Golding H. Expression and function of CCR5 and CXCR4 on human Langerhans cells and macrophages: implications for HIV primary infection. Nature medicine. 1997 Dec 1;3(12):1369-75.
2. Nduati R, John G, Mbori-Ngacha D, Richardson B, Overbaugh J, Mwatha A, Ndinya-Achola J, Bwayo J, Onyango FE, Hughes J, Kreiss J. Effect of breastfeeding and formula feeding on transmission of HIV-1: a randomized clinical trial. Jama. 2000 Mar 1;283(9):1167-74.
3. Ariën KK, Jespers V, Vanham G. HIV sexual transmission and microbicides. Reviews in medical virology. 2011 Mar;21(2):110-33.
4. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, Heldebrant C, Smith R, Conrad A, Kleinman SH, Busch MP. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. Aids. 2003 Sep 5;17(13):1871-9.
5. Fauci AS, Pantaleo G, Stanley S, Weissman D. Immunopathogenic mechanisms of HIV infection. Annals of internal medicine. 1996 Apr 1;124(7):654-63.
6. Eisinger RW, Dieffenbach CW, Fauci AS. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. Jama. 2019 Feb 5;321(5):451-2.
7. Rong L, Perelson AS. Modeling latently infected cell activation: viral and latent reservoir persistence, and viral blips in HIV-infected patients on potent therapy. PLoS Comput Biol. 2009 Oct 16;5(10):e1000533.
8. UNAIDS. 2020. Global HIV & AIDS statistics-2020 fact sheet. Joint United Nations Programme on HIV/AIDS (UNAIDS), Geneva, Switzerland.
9. Centers for Disease Control and Prevention. 2020. ABOUT HIV. Centers for Disease Control and Prevention, Atlanta, United States
10. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH. HIV-1 nomenclature proposal. Science. 2000 Apr 7;288(5463):55-.

11. Perrin L, Kaiser L, Yerly S. Travel and the spread of HIV-1 genetic variants. The Lancet infectious diseases. 2003 Jan 1;3(1):22-7.
12. Burke DS. Recombination in HIV: an important viral evolutionary strategy. Emerging infectious diseases. 1997 Jul;3(3):253.
13. Joseph SB, Swanstrom R, Kashuba AD, Cohen MS. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. Nature reviews Microbiology. 2015 Jul;13(7):414-25.
14. Arnold KB, Burgener A, Birse K, Romas L, Dunphy LJ, Shahabi K, Abou M, Westmacott GR, McCorrister S, Kwatampora J, Nyanga B. Increased levels of inflammatory cytokines in the female reproductive tract are associated with altered expression of proteases, mucosal barrier proteins, and an influx of HIV-susceptible target cells. Mucosal immunology. 2016 Jan;9(1):194-205.
15. Lumngwena EN, Metenou S, Masson L, Cicala C, Arthos J, Woodman Z. HIV-1 subtype C transmitted founders modulate dendritic cell inflammatory responses. Retrovirology. 2020 Dec;17(1):1-3.
16. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proceedings of the National Academy of Sciences. 2008 May 27;105(21):7552-7.
17. Parrish NF, Gao F, Li H, Giorgi EE, Barbian HJ, Parrish EH, Zajic L, Iyer SS, Decker JM, Kumar A, Hora B. Phenotypic properties of transmitted founder HIV-1. Proceedings of the National Academy of Sciences. 2013 Apr 23;110(17):6626-33.
18. Tully DC, Ogilvie CB, Batorsky RE, Bean DJ, Power KA, Ghebremichael M, Bedard HE, Gladden AD, Seese AM, Amero MA, Lane K. Differences in the selection bottleneck between modes of sexual transmission influence the genetic composition of the HIV-1 founder virus. PLoS pathogens. 2016 May 10;12(5):e1005619.
19. Go EP, Liao HX, Alam SM, Hua D, Haynes BF, Desaire H. Characterization of host-cell line specific glycosylation profiles of early transmitted/founder HIV-1

gp120 envelope proteins. Journal of proteome research. 2013 Mar 1;12(3):1223-34.

20. Chan DC, Fass D, Berger JM, Kim PS. Core structure of gp41 from the HIV envelope glycoprotein. Cell. 1997 Apr 18;89(2):263-73.
21. Wu L, Gerard NP, Wyatt R, Choe H, Parolin C, Ruffing N, Borsetti A, Cardoso AA, Desjardin E, Newman W, Gerard C. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. Nature. 1996 Nov;384(6605):179-83.
22. Hallenberger S, Bosch V, Angliker H, Shaw E, Klenk HD, Garten W. Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gpl60. Nature. 1992 Nov;360(6402):358-61.
23. Moulard M, Decroly E. Maturation of HIV envelope glycoprotein precursors by cellular endoproteases. Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes. 2000 Nov 10;1469(3):121-32.
24. Bonomelli C, Doores KJ, Dunlop DC, Thaney V, Dwek RA, Burton DR, Crispin M, Scanlan CN. The glycan shield of HIV is predominantly oligomannose independently of production system or viral clade. PloS one. 2011 Aug 16;6(8):e23521.
25. Zhu X, Borchers C, Bienstock RJ, Tomer KB. Mass spectrometric characterization of the glycosylation pattern of HIV-gp120 expressed in CHO cells. Biochemistry. 2000 Sep 19;39(37):11194-204.
26. Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, Darvill AG, Kinoshita T, Packer NH, Prestegard JH, Schnaar RL. Essentials of Glycobiology [internet].
27. Sanders RW, Venturi M, Schiffner L, Kalyanaraman R, Katinger H, Lloyd KO, Kwong PD, Moore JP. The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. Journal of virology. 2002 Jul 15;76(14):7293-305.
28. Julien JP, Sok D, Khayat R, Lee JH, Doores KJ, Walker LM, Ramos A, Diwanji DC, Pejchal R, Cupo A, Katpally U. Broadly neutralizing antibody PGT121

allosterically modulates CD4 binding via recognition of the HIV-1 gp120 V3 base and multiple surrounding glycans. PLoS Pathog. 2013 May 2;9(5):e1003342.

29. Sok D, Doores KJ, Briney B, Le KM, Saye-Francisco KL, Ramos A, Kulp DW, Julien JP, Menis S, Wickramasinghe L, Seaman MS. Promiscuous glycan site recognition by antibodies to the high-mannose patch of gp120 broadens neutralization of HIV. Science translational medicine. 2014 May 14;6(236):236ra63-.
30. Sattentau QJ. Envelope glycoprotein trimers as HIV-1 vaccine immunogens. Vaccines. 2013 Dec;1(4):497-512.
31. Doores KJ. The HIV glycan shield as a target for broadly neutralizing antibodies. The FEBS journal. 2015 Dec;282(24):4679-91.
32. Garces F, Sok D, Kong L, McBride R, Kim HJ, Saye-Francisco KF, Julien JP, Hua Y, Cupo A, Moore JP, Paulson JC. Structural evolution of glycan recognition by a family of potent HIV antibodies. Cell. 2014 Sep 25;159(1):69-79.
33. Ying H, Ji X, Hart ML, Gupta K, Saifuddin M, Zariffard MR, Spear GT. Interaction of mannose-binding lectin with HIV type 1 is sufficient for virus opsonization but not neutralization. AIDS research and human retroviruses. 2004 Mar 1;20(3):327-35.
34. Marzi A, Mitchell DA, Chaipan C, Fisch T, Doms RW, Carrington M, Desrosiers RC, Pöhlmann S. Modulation of HIV and SIV neutralization sensitivity by DC-SIGN and mannose-binding lectin. Virology. 2007 Nov 25;368(2):322-30.
35. Saifuddin M, Hart ML, Gewurz H, Zhang Y, Spear GT. Interaction of mannose-binding lectin with primary isolates of human immunodeficiency virus type 1. Journal of General Virology. 2000 Apr 1;81(4):949-55.
36. Silver ZA, Antonopoulos A, Haslam SM, Dell A, Dickinson GM, Seaman MS, Desrosiers RC. Discovery of O-Linked Carbohydrate on HIV-1 Envelope and Its Role in Shielding against One Category of Broadly Neutralizing Antibodies. Cell Reports. 2020 Feb 11;30(6):1862-9.
37. Raska M, Takahashi K, Czernekova L, Zachova K, Hall S, Moldoveanu Z, Elliott MC, Wilson L, Brown R, Jancova D, Barnes S. Glycosylation patterns of HIV-1

gp120 depend on the type of expressing cells and affect antibody recognition. Journal of Biological Chemistry. 2010 Jul 2;285(27):20860-9.

38. Termini JM, Church ES, Silver ZA, Haslam SM, Dell A, Desrosiers RC. HIV and SIV maintain high levels of infectivity in the complete absence of mucin type O-glycosylation. Journal of Virology. 2017 Jul 26.
39. Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology. 2004 Dec 1;14(12):1229-46.
40. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. British medical bulletin. 2001 Sep 1;58(1):19-42.
41. Doores KJ, Bonomelli C, Harvey DJ, Vasiljevic S, Dwek RA, Burton DR, Crispin M, Scanlan CN. Envelope glycans of immunodeficiency virions are almost entirely oligomannose antigens. Proceedings of the National Academy of Sciences. 2010 Aug 3;107(31):13800-5.
42. Curlin ME, Zioni R, Hawes SE, Liu Y, Deng W, Gottlieb GS, Zhu T, Mullins JI. HIV-1 envelope subregion length variation during disease progression. PLoS Pathog. 2010 Dec 16;6(12):e1001228.
43. Izquierdo-Useros N, Naranjo-Gómez M, Erkizia I, Puertas MC, Borràs FE, Blanco J, Martinez-Picado J. HIV and mature dendritic cells: Trojan exosomes riding the Trojan horse?. PLoS Pathog. 2010 Mar 26;6(3):e1000740.
44. Geijtenbeek TB, Kwon DS, Torensma R, van Vliet SJ, van Duijnhoven GC, Middel J, Cornelissen IL, Nottet HS, KewalRamani VN, Littman DR, Figdor CG. DC-SIGN, a dendritic cell–specific HIV-1-binding protein that enhances trans-infection of T cells. Cell. 2000 Mar 3;100(5):587-97.
45. Lambert AA, Gilbert C, Richard M, Beaulieu AD, Tremblay MJ. The C-type lectin surface receptor DCIR acts as a new attachment factor for HIV-1 in dendritic cells and contributes to trans-and cis-infection pathways. Blood. 2008 Aug 15;112(4):1299-307.

46. Izquierdo-Useros N, Lorizate M, Puertas MC, Rodriguez-Plata MT, Zangger N, Erikson E, Pino M, Erkizia I, Glass B, Clotet B, Keppler OT. Siglec-1 is a novel dendritic cell receptor that mediates HIV-1 trans-infection through recognition of viral membrane gangliosides. PLoS Biol. 2012 Dec 18;10(12):e1001448.
47. de Witte L, Nabatov A, Geijtenbeek TB. Distinct roles for DC-SIGN+-dendritic cells and Langerhans cells in HIV-1 transmission. Trends in molecular medicine. 2008 Jan 1;14(1):12-9.
48. Turville S, Wilkinson J, Cameron P, Dable J, Cunningham AL. The role of dendritic cell C-type lectin receptors in HIV pathogenesis. Journal of leukocyte biology. 2003 Nov;74(5):710-8.
49. Fahrbach KM, Barry SM, Ayehunie S, Lamore S, Klausner M, Hope TJ. Activated CD34-derived Langerhans cells mediate transinfection with human immunodeficiency virus. Journal of virology. 2007 Jul 1;81(13):6858-68.
50. Jiang AP, Jiang JF, Guo MG, Jin YM, Li YY, Wang JH. Human blood-circulating basophils capture HIV-1 and mediate viral trans-infection of CD4+ T cells. Journal of virology. 2015 Aug 1;89(15):8050-62.
51. Ping LH, Joseph SB, Anderson JA, Abrahams MR, Salazar-Gonzalez JF, Kincer LP, Treurnicht FK, Arney L, Ojeda S, Zhang M, Keys J. Comparison of viral Env proteins from acute and chronic infections with subtype C human immunodeficiency virus type 1 identifies differences in glycosylation and CCR5 utilization and suggests a new strategy for immunogen design. Journal of virology. 2013 Jul 1;87(13):7218-33.
52. Wilen CB, Parrish NF, Pfaff JM, Decker JM, Henning EA, Haim H, Petersen JE, Wojcechowskyj JA, Sodroski J, Haynes BF, Montefiori DC. Phenotypic and immunologic comparison of clade B transmitted/founder and chronic HIV-1 envelope glycoproteins. Journal of virology. 2011 Sep 1;85(17):8514-27.
53. Jiang C, Parrish NF, Wilen CB, Li H, Chen Y, Pavlicek JW, Berg A, Lu X, Song H, Tilton JC, Pfaff JM. Primary infection by a human immunodeficiency virus with atypical coreceptor tropism. Journal of virology. 2011 Oct 15;85(20):10669-81.

54. Schuitemaker H, Koot M, Kootstra NA, Dercksen MW, De Goede RE, Van Steenwijk RP, Lange JM, Schattenkerk JK, Miedema F, Tersmette M. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. Journal of virology. 1992 Mar 1;66(3):1354-60.
55. Parker ZF, Iyer SS, Wilen CB, Parrish NF, Chikere KC, Lee FH, Didigu CA, Berro R, Klasse PJ, Lee B, Moore JP. Transmitted/founder and chronic HIV-1 envelope proteins are distinguished by differential utilization of CCR5. Journal of virology. 2013 Mar 1;87(5):2401-11.
56. Del Portillo A, Tripodi J, Najfeld V, Wodarz D, Levy DN, Chen BK. Multiploid inheritance of HIV-1 during cell-to-cell infection. Journal of virology. 2011 Jul 15;85(14):7169-76.
57. Cardozo T, Kimura T, Philpott S, Weiser B, Burger H, Zolla-Pazner S. Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. AIDS research and human retroviruses. 2007 Mar 1;23(3):415-26.
58. Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM, Swanstrom R. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. Journal of virology. 2002 Apr 15;76(8):3852-64.
59. Clevestig P, Pramanik L, Leitner T, Ehrnst A. CCR5 use by human immunodeficiency virus type 1 is associated closely with the gp120 V3 loop N-linked glycosylation site. Journal of General Virology. 2006 Mar 1;87(3):607-12.
60. Sander O, Sing T, Sommer I, Low AJ, Cheung PK, Harrigan PR, Lengauer T, Domingues FS. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLoS Comput Biol. 2007 Mar 30;3(3):e58.
61. Balandya E, Sheth S, Sanders K, Wieland-Alter W, Lahey T. Semen protects CD4+ target cells from HIV infection but promotes the preferential transmission of R5 tropic HIV. The Journal of Immunology. 2010 Dec 15;185(12):7596-604.
62. Grivel JC, Shattock RJ, Margolis LB. Selective transmission of R5 HIV-1 variants: where is the gatekeeper?. Journal of translational medicine. 2011 Dec;9(1):1-7.

63. Cavarelli M, Foglieni C, Rescigno M, Scarlatti G. R5 HIV-1 envelope attracts dendritic cells to cross the human intestinal epithelium and sample luminal virions via engagement of the CCR5. EMBO molecular medicine. 2013 May;5(5):776-94.
64. Rhodes J. Investigating Mononuclear Phagocytes in Human Anogenital and Colorectal Tissues: Their Role in the Sexual Transmission of HIV (Doctoral dissertation, University of Sydney).
65. Rhodes JW, Tong O, Harman AN, Turville SG. Human dendritic cell subsets, ontogeny, and impact on HIV infection. Frontiers in immunology. 2019 May 16;10:1088.
66. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. AIDS (London, England). 2014 Jun 19;28(10):1509.
67. Klein K, Nickel G, Nankya I, Kyeyune F, Demers K, Ndashimye E, Kwok C, Chen PL, Rwambuya S, Poon A, Munjoma M. Higher sequence diversity in the vaginal tract than in blood at early HIV-1 infection. PLoS pathogens. 2018 Jan 18;14(1):e1006754.
68. Miller CJ, Li Q, Abel K, Kim EY, Ma ZM, Wietgrefe S, La Franco-Scheuch L, Compton L, Duan L, Shore MD, Zupancic M. Propagation and dissemination of infection after vaginal transmission of simian immunodeficiency virus. Journal of virology. 2005 Jul 1;79(14):9217-27.
69. Zhang ZQ, Schuler T, Zupancic M, Wietgrefe S, Staskus KA, Reimann KA, Reinhart TA, Rogan M, Cavert W, Miller CJ, Veazey RS. Sexual transmission and propagation of SIV and HIV in resting and activated CD4+ T cells. Science. 1999 Nov 12;286(5443):1353-7.
70. Petrova MI, van den Broek M, Balzarini J, Vanderleyden J, Lebeer S. Vaginal microbiota and its role in HIV transmission and infection. FEMS microbiology reviews. 2013 Sep 1;37(5):762-92.
71. Aldunate M, Tyssen D, Johnson A, Zakir T, Sonza S, Moench T, Cone R, Tachedjian G. Vaginal concentrations of lactic acid potently inactivate HIV. Journal of Antimicrobial Chemotherapy. 2013 Sep 1;68(9):2015-25.

72. Mitchell C, Hitti J, Paul K, Agnew K, Cohn SE, Luque AE, Coombs R. Cervicovaginal shedding of HIV type 1 is related to genital tract inflammation independent of changes in vaginal microbiota. AIDS research and human retroviruses. 2011 Jan 1;27(1):35-9.
73. Cone RA. Vaginal microbiota and sexually transmitted infections that may influence transmission of cell-associated HIV. The Journal of infectious diseases. 2014 Dec 15;210(suppl_3):S616-21.
74. Vujkovic-Cvijin I, Sortino O, Verheij E, Sklar J, Wit FW, Kootstra NA, Sellers B, Brenchley JM, Ananworanich J, van Der Loeff MS, Belkaid Y. HIV-associated gut dysbiosis is independent of sexual practice and correlates with noncommunicable diseases. Nature communications. 2020 May 15;11(1):1-5.
75. Nowak P, Troseid M, Avershina E, Barqasho B, Neogi U, Holm K, Hov JR, Noyan K, Vesterbacka J, Svärd J, Rudi K. Gut microbiota diversity predicts immune status in HIV-1 infection. Aids. 2015 Nov 28;29(18):2409-18.
76. Elopre L, Rodriguez M. Fecal microbiota therapy for recurrent Clostridium difficile infection in HIV-infected persons. Annals of internal medicine. 2013 May 21;158(10):779-80.
77. Vujkovic-Cvijin I, Rutishauser RL, Pao M, Hunt PW, Lynch SV, McCune JM, Somsouk M. Limited engraftment of donor microbiome via one-time fecal microbial transplantation in treated HIV-infected individuals. Gut microbes. 2017 Sep 3;8(5):440-50.
78. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. PLos med. 2005 Oct 25;2(11):e298.
79. Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, Chen MZ, Sewankambo NK. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. The Lancet. 2007 Feb 24;369(9562):657-66.
80. Dinh MH, Hirbod T, Kigozi G, Okocha EA, Cianci GC, Kong X, Prodger JL, Broliden K, Kaul R, Serwadda D, Wawer MJ. No difference in keratin thickness

between inner and outer foreskins from elective male circumcisions in Rakai, Uganda. PLoS One. 2012 Jul 18;7(7):e41271.

81. Jayathunge PH, McBride WJ, MacLaren D, Kaldor J, Vallely A, Turville S. Male circumcision and HIV transmission; what do we know?. The open AIDS journal. 2014;8:31.
82. Prodger JL. Defining Immune Correlates of HIV Susceptibility in the Foreskin (Doctoral dissertation).
83. Fahrbach KM, Barry SM, Anderson MR, Hope TJ. Enhanced cellular responses and environmental sampling within inner foreskin explants: implications for the foreskin's role in HIV transmission. Mucosal immunology. 2010 Jul;3(4):410-8.
84. Prodger JL, Gray RH, Shannon B, Shahabi K, Kong X, Grabowski K, Kigozi G, Nalugoda F, Serwadda D, Wawer MJ, Reynolds SJ. Chemokine levels in the penile coronal sulcus correlate with HIV-1 acquisition and are reduced by male circumcision in Rakai, Uganda. PLoS pathogens. 2016 Nov 29;12(11):e1006025.
85. Kim KA, Yolamanova M, Zirafi O, Roan NR, Staendker L, Forssmann WG, Burgener A, Dejucq-Rainsford N, Hahn BH, Shaw GM, Greene WC. Semen-mediated enhancement of HIV infection is donor-dependent and correlates with the levels of SEVI. Retrovirology. 2010 Dec;7(1):1-2.
86. Introini A, Boström S, Bradley F, Gibbs A, Glaessgen A, Tjernlund A, Broliden K. Seminal plasma induces inflammation and enhances HIV-1 replication in human cervical tissue explants. PLoS pathogens. 2017 May 19;13(5):e1006402.
87. Münch J, Rücker E, Ständker L, Adermann K, Goffinet C, Schindler M, Wildum S, Chinnadurai R, Rajan D, Specht A, Giménez-Gallego G. Semen-derived amyloid fibrils drastically enhance HIV infection. Cell. 2007 Dec 14;131(6):1059-71.
88. Thielens NM, Tacnet-Delorme P, Arlaud GJ. Interaction of C1q and mannan-binding lectin with viruses. Immunobiology. 2002 Jan 1;205(4-5):563-74.
89. Garred P, Madsen HO, Balslev U, Hofmann BO, Pedersen C, Gerstoft J, Svejgaard A. Susceptibility to HIV infection and progression of AIDS in relation to variant alleles of mannose-binding lectin. The Lancet. 1997 Jan 25;349(9047):236-40.

90. Bouwman LH, Roep BO, Roos A. Mannose-binding lectin: clinical implications for infection, transplantation, and autoimmunity. Human immunology. 2006 Apr 1;67(4-5):247-56.
91. Ji X, Gewurz H, Spear GT. Mannose binding lectin (MBL) and HIV. Molecular immunology. 2005 Feb 1;42(2):145-52.
92. Bermejo-Jambrina M, Eder J, Helgers LC, Hertoghs N, Nijmeijer BM, Stunnenberg M, Geijtenbeek TB. C-type lectin receptors in antiviral immunity and viral escape. Frontiers in immunology. 2018 Mar 26;9:590.
93. Mukherjee S, Zheng H, Derebe MG, Callenberg KM, Partch CL, Rollins D, Propheter DC, Rizo J, Grabe M, Jiang QX, Hooper LV. Antibacterial membrane attack by a pore-forming intestinal C-type lectin. Nature. 2014 Jan;505(7481):103-7.
94. Breitenbach Barroso Coelho LC, Marcelino dos Santos Silva P, Felix de Oliveira W, De Moura MC, Viana Pontual E, Soares Gomes F, Guedes Paiva PM, Napoleão TH, dos Santos Correia MT. Lectins as antimicrobial agents. Journal of applied microbiology. 2018 Nov;125(5):1238-52.
95. da Silva JD, da Silva SP, da Silva PM, Vieira AM, de Araújo LC, de Albuquerque Lima T, de Oliveira AP, do Nascimento Carvalho LV, da Rocha Pitta MG, de Melo Rêgo MJ, Pinheiro IO. Portulaca elatior root contains a trehalose-binding lectin with antibacterial and antifungal activities. International journal of biological macromolecules. 2019 Apr 1;126:291-7.
96. Akkouh O, Ng TB, Singh SS, Yin C, Dan X, Chan YS, Pan W, Cheung RC. Lectins with anti-HIV activity: a review. Molecules. 2015 Jan;20(1):648-68.
97. Brenchley JM, Hill BJ, Ambrozak DR, Price DA, Guenaga FJ, Casazza JP, Kuruppu J, Yazdani J, Migueles SA, Connors M, Roederer M. T-cell subsets that harbor human immunodeficiency virus (HIV) in vivo: implications for HIV pathogenesis. Journal of virology. 2004 Feb 1;78(3):1160-8.
98. Heeregrave EJ, Geels MJ, Brenchley JM, Baan E, Ambrozak DR, van der Sluis RM, Bennemeer R, Douek DC, Goudsmit J, Pollakis G, Koup RA. Lack of in vivo compartmentalization among HIV-1 infected naive and memory CD4+ T cell subsets. Virology. 2009 Oct 10;393(1):24-32.

99. Bleul CC, Wu L, Hoxie JA, Springer TA, Mackay CR. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. Proceedings of the National Academy of Sciences. 1997 Mar 4;94(5):1925-30.

100. Okoye AA, Picker LJ. CD 4+ T-cell depletion in HIV infection: mechanisms of immunological failure. Immunological reviews. 2013 Jul;254(1):54-64.

101. McCune JM. The dynamics of CD4+ T-cell depletion in HIV disease. Nature. 2001 Apr;410(6831):974-9.

102. Tjomsland V, Ellegård R, Che K, Hinkula J, Lifson JD, Larsson M. Complement opsonization of HIV-1 enhances the uptake by dendritic cells and involves the endocytic lectin and integrin receptor families. PloS one. 2011 Aug 11;6(8):e23542.

103. Scarlatti G, Tresoldi E, Björndal Å, Fredriksson R, Colognesi C, Deng HK, Malnati MS, Plebani A, Siccardi AG, Littman DR, Fenyö EM. In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. Nature medicine. 1997 Nov;3(11):1259-65.

104. Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR. Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. The Journal of experimental medicine. 1997 Feb 17;185(4):621-8.

105. Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, Sodroski JG. The antigenic structure of the HIV gp120 envelope glycoprotein. Nature. 1998 Jun;393(6686):705-11.

106. Zhang W, Canziani G, Plugariu C, Wyatt R, Sodroski J, Sweet R, Kwong P, Hendrickson W, Chaiken I. Conformational changes of gp120 in epitopes near the CCR5 binding site are induced by CD4 and a CD4 miniprotein mimetic. Biochemistry. 1999 Jul 20;38(29):9405-16.

107. Chan DC, Fass D, Berger JM, Kim PS. Core structure of gp41 from the HIV envelope glycoprotein. Cell. 1997 Apr 18;89(2):263-73.

108. Rinaldo CR. HIV-1 trans infection of CD4+ T cells by professional antigen presenting cells. Scientifica. 2013 Oct;2013.

109. Bouhlal H, Chomont N, Réquena M, Nasreddine N, Saidi H, Legoff J, Kazatchkine MD, Bélec L, Hocini H. Opsonization of HIV with complement enhances infection of dendritic cells and viral transfer to CD4 T cells in a CR3 and DC-SIGN-dependent manner. The Journal of Immunology. 2007 Jan 15;178(2):1086-95.

110. Jolly C, Sattentau QJ. Retroviral spread by induction of virological synapses. Traffic. 2004 Sep;5(9):643-50.

111. de Witte L, Nabatov A, Pion M, Fluitsma D, De Jong MA, de Gruijl T, Piguet V, van Kooyk Y, Geijtenbeek TB. Langerin is a natural barrier to HIV-1 transmission by Langerhans cells. Nature medicine. 2007 Mar;13(3):367-71.

112. Peressin M, Proust A, Schmidt S, Su B, Lambotin M, Biedma ME, Laumond G, Decoville T, Holl V, Moog C. Efficient transfer of HIV-1 in trans and in cis from Langerhans dendritic cells and macrophages to autologous T lymphocytes. Aids. 2014 Mar 13;28(5):667-77.

113. Bobardt MD, Chatterji U, Selvarajah S, Van der Schueren B, David G, Kahn B, Gallay PA. Cell-free human immunodeficiency virus type 1 transcytosis through primary genital epithelial cells. Journal of virology. 2007 Jan 1;81(1):395-405.

114. Carias AM, McCoombe S, McRaven M, Anderson M, Galloway N, Vandergrift N, Fought AJ, Lurain J, Duplantis M, Veazey RS, Hope TJ. Defining the interaction of HIV-1 with the mucosal barriers of the female reproductive tract. Journal of virology. 2013 Nov 1;87(21):11388-400.

115. Kinlock BL, Wang Y, Turner TM, Wang C, Liu B. Transcytosis of HIV-1 through vaginal epithelial cells is dependent on trafficking to the endocytic recycling pathway. PloS one. 2014 May 15;9(5):e96760.

116. Moyes DL, Islam A, Kohli A, Naglik JR. Oral epithelial cells and their interactions with HIV-1. Oral diseases. 2016 Apr;22:66-72.

117. Dudley DM, Gao Y, Nelson KN, Henry KR, Nankya I, Gibson RM, Arts EJ. A novel yeast-based recombination method to clone and propagate diverse HIV-1 isolates. Biotechniques. 2009 May;46(6):458-67.

118. Princen K, Hatse S, Vermeire K, De Clercq E, Schols D. Establishment of a novel CCR5 and CXCR4 expressing CD4+ cell line which is highly sensitive to HIV and suitable for high-throughput evaluation of CCR5 and CXCR4 antagonists. Retrovirology. 2004 Dec;1(1):1-3.

119. Ordoño D, Enjuanes L, Casasnovas JM. Methods for preparation of low abundance glycoproteins from mammalian cell supernatants. International journal of biological macromolecules. 2006 Aug 15;39(1-3):151-6.

120. Cao L, Pauthner M, Andrabi R, Rantalainen K, Berndsen Z, Diedrich JK, Menis S, Sok D, Bastidas R, Park SK, Delahunty CM. Differential processing of HIV envelope glycans on the virus and soluble recombinant trimer. Nature communications. 2018 Sep 12;9(1):1-4.

121. Barbouche R, Miquelis R, Jones IM, Fenouillet E. Protein-disulfide isomerase-mediated reduction of two disulfide bonds of HIV envelope glycoprotein 120 occurs post-CXCR4 binding and is required for fusion. Journal of Biological Chemistry. 2003 Jan 31;278(5):3131-6.

122. Moore JP, McKeating JA, Weiss RA, Sattentau QJ. Dissociation of gp120 from HIV-1 virions induced by soluble CD4. Science. 1990 Nov 23;250(4984):1139-42.

123. Lu M, Ma X, Reichard N, Terry DS, Arthos J, Smith AB, Sodroski JG, Blanchard SC, Mothes W. Shedding-resistant HIV-1 envelope glycoproteins adopt downstream conformations that remain responsive to conformation-preferring ligands. Journal of Virology. 2020 Aug 17;94(17).

124. Yang L, Song Y, Li X, Huang X, Liu J, Ding H, Zhu P, Zhou P. HIV-1 virus-like particles produced by stably transfected Drosophila S2 cells: a desirable vaccine component. Journal of virology. 2012 Jul 15;86(14):7662-76.

125. Cao L, Pauthner M, Andrabi R, Rantalainen K, Berndsen Z, Diedrich JK, Menis S, Sok D, Bastidas R, Park SK, Delahunty CM. Differential processing of HIV envelope glycans on the virus and soluble recombinant trimer. Nature communications. 2018 Sep 12;9(1):1-4

126. Struwe WB, Chertova E, Allen JD, Seabright GE, Watanabe Y, Harvey DJ, Medina-Ramirez M, Roser JD, Smith R, Westcott D, Keele BF. Site-specific

glycosylation of virion-derived HIV-1 Env is mimicked by a soluble trimeric immunogen. Cell reports. 2018 Aug 21;24(8):1958-66.

127. GILLJAM G. Envelope glycoproteins of HIV-1, HIV-2, and SIV purified with Galanthus nivalis agglutinin induce strong immune responses. AIDS research and human retroviruses. 1993 May;9(5):431-8

128. Sok D, Doores KJ, Briney B, Le KM, Saye-Francisco KL, Ramos A, Kulp DW, Julien JP, Menis S, Wickramasinghe L, Seaman MS. Promiscuous glycan site recognition by antibodies to the high-mannose patch of gp120 broadens neutralization of HIV. Science translational medicine. 2014 May 14;6(236):236ra63-.

129. Liu G, Cheng K, Lo CY, Li J, Qu J, Neelamegham S. A comprehensive, open-source platform for mass spectrometry-based glycoproteomics data analysis. Molecular & Cellular Proteomics. 2017 Nov 1;16(11):2032-47.

Appendices

Appendix A: MS settings for Orbitrap Fusion Lumos

Global Settings

Use Ion Source Settings from Tune = False

Method Duration (min)= 60

Ion Source Type = NSI

Spray Voltage: Positive Ion (V) = 1900

Spray Voltage: Negative Ion (V) = 600

Sweep Gas (Arb) = 0

Ion Transfer Tube Temp (∞C) = 250

APPI Lamp = Not in use

Internal Mass Calibration= User Defined Lock Mass

Pressure Mode = Standard

Default Charge State = 2

Advanced Precursor Determination = False

Internal Cal Positive

m/z

445.12003

Start Time (min) = 0

End Time (min) = 60

Cycle Time (sec) = 3

Scan MasterScan

MSn Level = 1

Use Wide Quad Isolation = True

Detector Type = Orbitrap

Orbitrap Resolution = 120K

Mass Range = Normal

Scan Range (m/z) = 350-1800

Maximum Injection Time (ms) = 50

AGC Target = 400000

Microscans = 1

RF Lens (%) = 30

Use ETD Internal Calibration = False

DataType = Profile

Polarity = Positive

Source Fragmentation = False

Scan Description =

Filter ChargeState

Include undetermined charge states = False

Include charge state(s) = 2-8

Include charge states 25 and higher = False

Filter DynamicExclusion

Exclude after n times = 1

Exclusion duration (s) = 15

Mass Tolerance = ppm

Mass tolerance low = 10

Mass tolerance high = 10

Exclude isotopes = True

Perform dependent scan on single charge state per precursor only = False

Filter IntensityThreshold

Intensity Filter Type = IntensityThreshold

Maximum Intensity = 1E+20

Minimum Intensity = 50000

Relative Intensity Threshold = 0

Data Dependent Properties

Data Dependent Mode= Cycle Time

Scan Event 1

Filter PrecursorPriority

HighestChargeState

Filter PrecursorPriority

LowestM/Z

Scan ddMSnScan

MSn Level = 2

Isolation Mode = Quadrupole

Isolation Offset = Off

Isolation Window = 2

Reported Mass = Offset Mass

Multi-notch Isolation = False

Scan Range Mode = Auto Normal

FirstMass = 120

Scan Priority= 1

ActivationType = HCD

Is Stepped Collision Energy On = False

Stepped Collision Energy (%) = 5

Collision Energy (%) = 28

Detector Type = Orbitrap

Orbitrap Resolution = 30K

Maximum Injection Time (ms) = 60

AGC Target = 50000

Inject ions for all available parallelizable time = False

Microscans = 1

Use ETD Internal Calibration = False

DataType = Profile

Polarity = Positive

Source Fragmentation = False

Scan Description =

Filter ProductIonTrigger

trigger only when at least n product ions from list are detected = True

n = : 1

product ion(s) must be within top n = True

Top N Product Ions = 20

Trigger Only When Correct Charge State of Product Ion is Detected = True

Ignore charge state requirement for unassigned ions = True

product ions(s) must be above threshold (relative intensity, %) = False

Product Ions Threshold = 0

>>>>>>>>>>>>> Mass List Table <<<<<<<<<<<<<

CompoundName	m/z
HexNAc	204.0867
HexNAcfragment	138.0545
HexNAcHex	366.1396

Data Dependent Properties

Data Dependent Mode= Number of Scans

Number of Dependent Scans= 2

Scan ddMSnScan

MSn Level = 2

Isolation Mode = Quadrupole

Isolation Offset = Off

Isolation Window = 3

Reported Mass = Offset Mass

Multi-notch Isolation = False

Scan Range Mode = Define m/z range

Scan Priority= 1

ActivationType = ETD

Is EThcD Active = True

Use calibrated charge dependent ETD parameters = True

Detector Type = Orbitrap

Orbitrap Resolution = 30K

Scan Range (m/z) = 120-2000

Maximum Injection Time (ms) = 250

AGC Target = 200000

Inject ions for all available parallelizable time = False

Microscans = 1

Use ETD Internal Calibration = False

DataType = Profile

Polarity = Positive

Source Fragmentation = False

Scan Description =

Scan ddMSnScan

MSn Level = 2

Isolation Mode = Quadrupole

Isolation Offset = Off

Isolation Window = 3

Reported Mass = Offset Mass

Multi-notch Isolation = False

Scan Range Mode = Define m/z range

Scan Priority= 1

ActivationType = ETD

Is EThcD Active = True

Use calibrated charge dependent ETD parameters = True

Detector Type = Orbitrap

Orbitrap Resolution = 30K

Scan Range (m/z) = 120-2000

Maximum Injection Time (ms) = 250

AGC Target = 200000

Inject ions for all available parallelizable time = False

Microscans = 1

Use ETD Internal Calibration = False

DataType = Profile

Polarity = Positive

Source Fragmentation = False

Appendix B: Protocol for Initial Computerized Search using GlycoPAT in Matlab

Disclaimer: The use of GlycoPAT to analyze the MS data was originally published in 2014 and employed by Dr. Najwa Zebian from the Creuzenet lab in 2018 in the Linux system[99]. The document presented below describes an improved protocol that works in both Windows and Linux based on the initial setup.

Introduction: After receiving the raw MS data from the Toronto SPARC facility, a 3-step data analysis process begins. This document describes the first step: the initial computerized search. Each raw dataset contains tens of thousands of spectra each containing peaks representing ions. The mode of MS used in this study is ETHcD, a combination of HCD and ETD. HCD produces B and Y type ions, and ETD produces C and Z type ions. Since spectra cannot be readily understood in context of peptides and glycans, the goal of the initial computerized search is to translate the spectra from picture to a peptide or glycopeptide assignment, which is a peptide sequence potentially carrying glycans noted in standardized glycobiology nomenclature. This is achieved by matching the m/z values of the ions from the experimental datasets to a theoretical library of m/z values of all possible ions that can be generated from the given protein sequence, pool of potential post-translational modifications (PTMs), and enzymes used to digest the full-length gp120. The calculation of the theoretical library and the subsequent matching between the experimental m/z values to the library are both performed using the GlycoPAT software in the Matlab environment. After the matching process is completed for each dataset, an output csv file containing the peptide and glycopeptide assignments translated from the matched spectra is generated. The second step of the MS data analysis is to organize the output csv file and extract the biological meaning of the data, as described in detail in appendix C. The third step is to evaluate the statistical significance of the differences in gp120 glycosylation found between the two viral strains.

1. Install Matlab by first downloading the installation file from https://www.mathworks.com/products/matlab.html. Then install the parallel computing toolbox in Matlab. Open Matlab, go to Home>get Add-Ons, search for Parallel computing toolbox, and click install.

2. Install GlycoPAT. For detailed instructions, visit pages 3 and 4 of GlycoPAT manual, downloadable at: https://sourceforge.net/projects/glycopat/files/GlycoPAT_Manual.pdf/download.

3. Download the raw MS data files from the link provided by the MS facility. Move the downloaded files into a user-specific folder, e.g. “gp120 project”. The raw data files are in the format of “.raw”.

4. Convert the raw files into the MzXML format using a software known as MSConvert from Proteowizard, downloadable at: http://proteowizard.sourceforge.net/download.html.

5. Create a folder for each MzXML file to prepare to run the initial computerized search. In the same folder, there are 6 required program files in addition to the MzXML file (sample folder downloadable at https://drive.google.com/drive/folders/129IR7weOydROSpvnsDbzkMxUPq9DPaRX?usp=sharing):

    a. Target protein sequence in fasta format

        i. In this study, B4 and Q0 have different gp120 sequences. Check that the right fasta file is used for the right MzXML dataset.

        ii. To edit the fasta sequence: change the file extension from .fasta to .txt to open file in notepad and make the changes. Save and close notepad. Change the file extension from .txt back to .fasta.

    b. Variable_modifications_custom.txt (contains a list of PTMs)

        i. This list currently contains 29 glycoforms and 2 PTMs (oxidation and carbamidomethylation) introduced during protein band processing.

        ii. This list may be modified manually if the user is interested in adding or reducing PTMs. When adding glycoforms, check that

the glycoform codes match the GlycoPAT nomenclature (details in GlycoPAT initial publication: A Comprehensive, Open-source Platform for Mass Spectrometry-based Glycoproteomics Data Analysis)

c. Fixed_modifications_2.txt (no content but required for the program to run)

d. scoreAll.mako (program file, do not modify)

e. gather.txt (program file, do not modify)

f. batch.m (the script that contains instructions for the run and must be edited specifically for each run)

6. Each run requires an individual folder. Copy all necessary files from the sample folder to the folder containing the MzXML data file. There are two versions of the batch.m file: long and short. The long version contains instructions for both steps of generating the theoretical library (named digestedpep.txt) and matching the experimental values to the library. Once the library is generated, it does not need to be generated again for other datasets of the same protein sequence, glycoforms/PTMs, and enzymes used. Therefore, the short version, which only contains instructions for the second matching step, can be used with the pre-generated library. In this case, simply copy the pre-generated library from the previous run to the folder of the dataset to be run.
7. When all necessary files are in one folder, check that the fasta sequence and list of glycoforms/PTMs are correct and the appropriate version of batch.m is chosen.

8. Double click the batch.m file to open it in Matlab. Below is a screenshot of the long version of batch.m. Edit the fields highlighted in red, make sure to edit within the single quotation marks. The field in line 14 is the name of the fasta file. Line 28 is the path of the folder the MzXML file is in. Line 29 is the name of the MzXML file. The short version of batch.m does not have lines 14 to 22, which are the codes for the first step of generating the digestedpep.txt library. Edit these fields as necessary by directly overwriting the code boxes in red.

```
12
13      % configure protein digestion analysis
14 -    fasta = peptideread('gp120_Q0.fasta');
15 -    fixmod = fixedptmread('fixed_modifications_2.txt');
16 -    varmod = varptmread('variable_modifications_custom.txt');
17 -    options = digestoptionset('missedmax', 4, 'minpeplen', 4, 'maxpeplen', 25, 'minptm', 0, 'maxptm', 4);
18 -    options = digestoptionset(options, 'fixedptm', fixmod, 'varptm', varmod);
19 -    options = digestoptionset(options, 'isoutputfile', 'yes');
20      % default output file name is "digestedpep.txt"
21
22 -    fragments = digestSGP(fasta, {'Trypsin', 'Chymotrypsin Low'}, options);
23      % disp(fragments);
24
25      % Tandem MS Analysis settings
26 -    pepfile = 'digestedpep.txt';  % output from last step
27      %xmlfilepath = '/home/art/work/zebian/';
28 -    xmlfilepath = 'C:\Users\creuz\Desktop\YS_gp120\corrected batch\Q0_4_ELU_80\';
29 -    mzXMLfilename = 'Q0_4_ELU_80.mzXML';
30 -    fragMode = 'AUTO';
31 -    MS1tol = 10.0;
32 -    MS1tolUnit = 'ppm';
33 -    MS2tol = 10.0;
34 -    MS2tolUnit = 'ppm';
35 -    outputDir = xmlfilepath;
36 -    outputCSV = 'tandemMS.csv';
37 -    maxlag = 50;
38 -    cutoffMed = 2.0;
39 -    fracMax = 0.02;
40 -    nmFrag = 0;
41 -    npFrag = 2;
42 -    ngFrag = 0;
43 -    selectPeak = [204.087, 163.061, 147.066, 310.278, 292.267, 366.14, 407.147];
```

9. After editing the batch.m, click the green “Run” button in the commend bar on top, and the green “Run” button will become a blue “Pause” button. Do not click on the “Pause” button. If the long version is executed, the run typically takes around a week. If the short version is executed with the pre-generated library file, the run typically takes less than 24 hours. Generating the library is the rate-limiting step and the matching step is much faster.
10. When the program is running, there is no progress bar. As long as there is no error messages or warnings while the blue “Pause” button is seen, the program is running. When the run is finished, an output file named “tandomMS.csv” appears in the same folder and Matlab terminates itself.
11. This is the end of this protocol. The next step in MS data analysis begin with analyzing the contents of the “tandomMS.csv” output, which is described in detail in Appendix C.

Appendix C: Protocol for Analysis of GlycoPAT Output (csv file)

As outlined in the protocol of the previous step (Appendix B: Initial Search Using GlycoPAT in Matlab), raw data files were converted into mzMXL format and fed to the GlycoPAT software in Matlab. The output (TandomMS.csv) containing linear text interpretations (known as “smallglypep” in software, or sgp) of each scanned spectrum of the MS experiment can then be opened and manually analyzed in GlycoPAT and Excel. Since the load of information in the csv files were beyond human processing capacity, this protocol was specifically developed for this project as a semi-automated platform based in Excel. Using this method, each MS dataset was organized into a matrix of glycosylation sites and glycoform (GF) frequency at each site. Based on the original strain and band molecular weight from which the raw data were generated, the matrices were then compared and contrasted to characterize the gp120 glycosylation fingerprints of acute transmitted founder viruses and chronic viruses.

12. Rename the newly created TandomMS.csv file to the name of the dataset, eg. Q0_4_ELU_80. Save as excel workbook (.xlsx) and proceed to the following steps in the .xlsx file.

13. Open another excel workbook named “empty work station-make copy before use”. This workbook contains brief instructions and pre-entered formula. Copy the sheets in this workbook to yourdata.xlsx. Certain modifications must be made if the fasta sequence is not B4 or Q0.

14. Copy scan and sgp (columns B and G from step 1) to columns A and B of sheet “1.posi #, remove dups, pretreat”. Cell C2 has previously entered formula: =A2&" "&B2. This formula merges scan and sgp. Double click the bottom right corner of cell C2 to autofill down the entire list. Merging the scan # and sgp will ensure they will stay matched during the next steps.

15. Cleaning 1: create a new blank sheet. Copy merged column C to blank sheet and paste as values only. Remove <i> and <o> in sgp by replacing <*> with blank. <i> and <o> are carbamidomethylation and oxidation, respectively. These are

modifications artificially introduced during sample preparation and may alter the m/z values of ions. These modifications will be useful for individual spectrum annotation, but not useful for generating site-GF table. Some sgp differ by <*>. After removing <*>, there will be redundant sgps, which can be removed by clicking DATA>Remove duplicates to leave only unique sgp-scan pairs in each cell. The removed <*> information can be traced back in the original csv file according to the scan number, if necessary.

16. Cleaning 2: Apply filter to select rows that contain { or }. GlycoPAT annotates glycans within {}. This step selects sgps with glycans. *Sgps without glycans can be traced back by un-applying the filter.

17. Copy filtered list to new blank sheet. This list now comprises exclusively individual glycopeptides and associated scan numbers. Unmerge scan and sgp back to two columns using DATA>Text to columns. Choose “delimited”, next, check “space”, next, finish.

18. Copy unmerged scan and sgp columns back to columns E and F in sheet “1.posi #, remove dups, pretreat” to prepare the calculation of position number, the position of the first glycosylated N of the sgp within the entire fasta sequence, which can be used to sort the sgps in order of N to C termini.

    a. First find the position number of the **first glycosylated N**-site in the **sgp** using column G. Cell G2 has previously entered formula: =FIND("N{",F2). For example, if cell F2 contains MKN{n{f}{n}}CSF, cell G2 returns 3 because N{ is found starting at the third position in the text string. Double click the bottom right corner of cell G2 to autofill down the entire list.

    b. Then find the position number of the **first amino acid of the sgp** within the entire **fasta sequence** (column K). Glycans in sgps must be removed in order to use the FIND function to search the naked peptide in the fasta sequence. To do this, go back to the sheet in step 6, replace {*} with

nothing and then replace } with nothing. Copy this naked sgp list to column I of sheet “1.posi #, remove dups, pretreat”. Then copy the full fasta sequence of parent protein into cell J2. Double click the bottom right corner of cell J2 to autofill down the entire column. Cell K2 has previously entered formula: =FIND(I2,J2). This formula returns the starting position of the sgp peptide (cell I2) in the fasta sequence (cell J2) in cell K2. Double click the bottom right corner of cell K2 to autofill down the entire list.

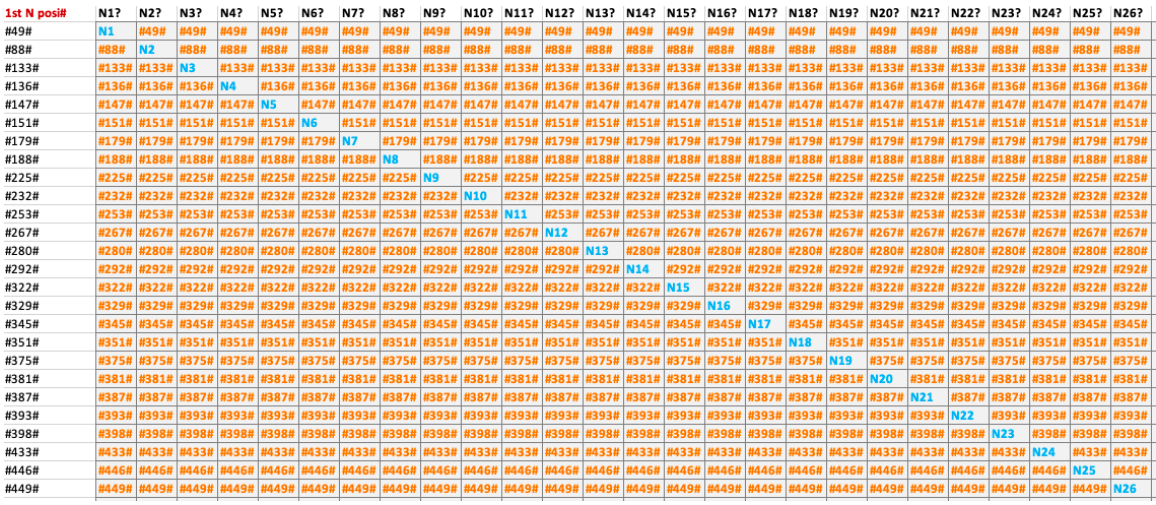
c. To obtain the position number of the **first glycosylated N of the sgp** within the **entire fasta sequence**, add the numbers from steps 7a and 7b in **column L**. Cell L2 sums the numbers in cell G2 and K2 using formula: =G2+K2. Double click the bottom right corner of cell L2 to autofill down the entire column. This number is then used to sort all the sgps such that all the sgps with the same N-site are ordered and grouped.

19. Copy scan list, 1st N posi #, and sgp (columns E, L, F) into columns N, O, Q, respectively. Select columns N, O, and Q, and sort by 1st N position number (column O) from smallest value to the largest. Do not sort any columns again until final verification step.

20. Now the list of 1st N position values in column O are sorted, they can be pretreated using pre-entered formulas (P2="#"&O2&"#") in column P. For example, if cell O2 contains position number 49, cell P2 shows #49#. Double click the bottom right corner of cell O2 to autofill the entire column P. Then copy column P to column A of sheet named “N posi# to Nx”. This sheet contains formulas that convert the residue position numbers of N-sites into simpler names for recognition and operation, such as N1, N2, N3, so on. The pretreatment step is necessary in order to avoid confusion when position numbers are replaced with Nx notation. For example, for the Q0 sequence, position number 49 is equivalent to N1 and position number 449 to N26. Since 449 contains 49, without

pretreatment, it becomes “4N1” instead of “N26”. With pretreatment, #49# and #449# always get converted properly to N1 and N26.

a. In the sheet named “N posi# to Nx”, column A now has the list of pretreated position numbers. Column B has pre-entered formulas that convert position numbers into Nx notation. The formula used is: B2=SUBSTITUTE(A2,"#49#","N1"). If A2 contains "#49#", B2 will show N1. Similarly, in cell C2, the formula will be C2=SUBSTITUTE(A2,"#88#","N2") and will display N2 if A2 contains "#88#". The picture below shows the schematics of how this sheet works based on the list of position numbers of Q0. B4 has a different list of position numbers due to different protein sequence.



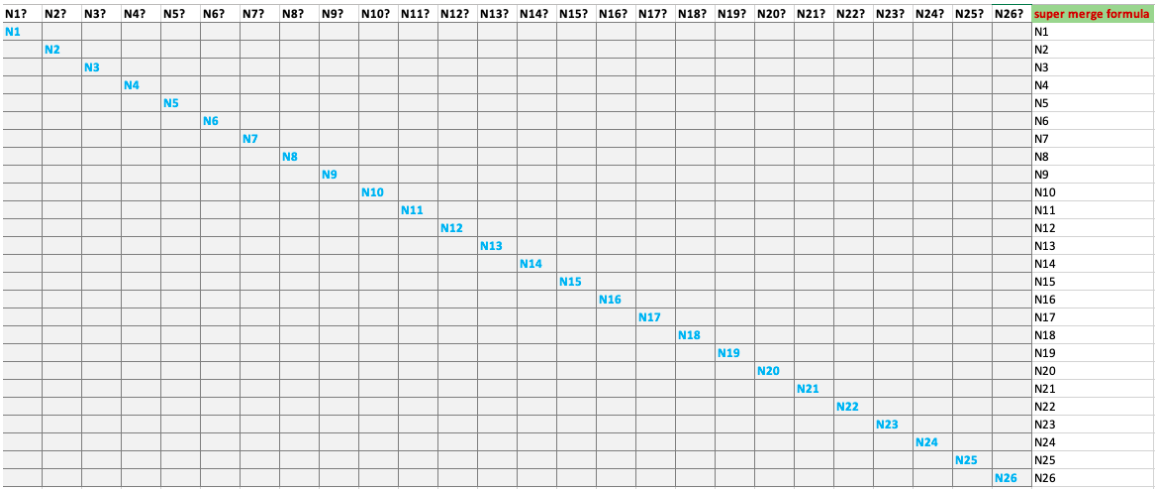
1st N posi#	N1?	N2?	N3?	N4?	N5?	N6?	N7?	N8?	N9?	N10?	N11?	N12?	N13?	N14?	N15?	N16?	N17?	N18?	N19?	N20?	N21?	N22?	N23?	N24?	N25?	N26?
#49#	N1	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#	#49#
#88#	#88#	N2	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#	#88#
#133#	#133#	#133#	N3	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#	#133#
#136#	#136#	#136#	#136#	N4	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#	#136#
#147#	#147#	#147#	#147#	#147#	N5	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#	#147#
#151#	#151#	#151#	#151#	#151#	#151#	N6	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#	#151#
#179#	#179#	#179#	#179#	#179#	#179#	#179#	N7	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#	#179#
#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	N8	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#	#188#
#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	N9	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#	#225#
#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	N10	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#	#232#
#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	N11	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#	#253#
#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	N12	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#	#267#
#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	N13	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#	#280#
#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	N14	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#	#292#
#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	N15	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#	#322#
#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	N16	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#	#329#
#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	N17	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#	#345#
#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#	N18	#351#	#351#	#351#	#351#	#351#	#351#	#351#	#351#
#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	#375#	N19	#375#	#375#	#375#	#375#	#375#	#375#	#375#
#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	#381#	N20	#381#	#381#	#381#	#381#	#381#	#381#
#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	#387#	N21	#387#	#387#	#387#	#387#	#387#
#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	#393#	N22	#393#	#393#	#393#	#393#
#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	#398#	N23	#398#	#398#	#398#
#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	#433#	N24	#433#	#433#
#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	#446#	N25	#446#
#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	#449#	N26

b. In case of Q0, there are 26 potential N-linked sites. Formulas in columns B to AA will automatically show N1 to N26 based on the input from column A.

c. Copy the calculation area and paste values only to a new area (replacement area, columns AD though BC for Q0 datasets). Then remove any cells that was not converted to the Nx format by highlighting the newly pasted area and replacing “#*#” with nothing.

d. The replacement area now looks largely empty and each row only has one cell containing Nx, which can be dispersed in any column in the

replacement area. To summarize the dispersed Nx cells in one column, use the "super-merge" function (Column BD, in case of Q0): BD2=AD2&AE2&AF2&AG2&AH2&AI2&AJ2&AK2&AL2&AM2&AN2&AO2&AP2&AQ2&AR2&AS2&AT2&AU2&AV2&AW2&AX2&AY2&AZ2&BA2&BB2&BC2. Double click the bottom right corner of cell BE2 to autofill down the entire list. This is shown in the picture below:



N1?	N2?	N3?	N4?	N5?	N6?	N7?	N8?	N9?	N10?	N11?	N12?	N13?	N14?	N15?	N16?	N17?	N18?	N19?	N20?	N21?	N22?	N23?	N24?	N25?	N26?	super merge formula
N1																										N1
	N2																									N2
		N3																								N3
			N4																							N4
				N5																						N5
					N6																					N6
						N7																				N7
							N8																			N8
								N9																		N9
									N10																	N10
										N11																N11
											N12															N12
												N13														N13
													N14													N14
														N15												N15
															N16											N16
																N17										N17
																	N18									N18
																		N19								N19
																			N20							N20
																				N21						N21
																					N22					N22
																						N23				N23
																							N24			N24
																								N25		N25
																									N26	N26

e. Copy super-merge column, paste values only to column named "1st N site" of sheet named "final table".

f. Remove "N" in the "1st N site" column to only keep numbers by highlighting this column and replacing N with nothing.

g. Recall this column represents the first occupied N site in the sgp and any potential second and third N sites in the sgp are not accounted for. The second N-site is simply the first N-site plus 1, and the third N-site is first N-site plus 2. In the columns named "2nd N site" and "3rd N site", formulas are pre-entered to perform such simple addition based on the "1st N site" column. Double click the bottom right corner of first cell in "2nd N site" to autofill the list. Repeat for "3rd N site". A schematic calculation of the 2nd and 3rd N-site is shown in the picture below:

Nx super merge	2nd N site	3rd N site
1	2	3
2	3	4
3	4	5
4	5	6
5	6	7
6	7	8
7	8	9
8	9	10
9	10	11
10	11	12
11	12	13
12	13	14
13	14	15
14	15	16
15	16	17
16	17	18
17	18	19
18	19	20
19	20	21
20	21	22
21	22	23
22	23	24
23	24	25
24	25	26
25	26	27
26	27	28

h. As we are only interested in glycosylated N-sites, step 9g poses a problem because it has two assumptions: 1, all sgp have multiple N-sites; 2, that all these N-sites are glycosylated. Assumption 1 does not affect later steps but assumption 2 could results in inaccuracy of the final table. A verification step is performed to resolve this problem, explained in detail later.

21. Now N-sites are converted to numerical format, glycoforms must also be converted to a short numerical format for further analysis. Go back to sheet named “posi #, remove dups, pretreat”. Pretreat the sgp in column Q to eventually convert glycoforms from “{}” notion into “GF1”, “GF2”, so on.

    a. Replace “N{” with “N {”. Make sure to check mark “match case” option before replacement. For example, the sgp “AN{n{f}{n{h}}}TTLF” becomes “AN {n{f}{n{h}}}TTLF”.

    b. Then copy pretreated sgp to column S. Go to DATA>Text to columns. Choose “delimited”, next, check “space”. This action splits cell contents in column S into multiple cells through columns S to V using the spaces created in step 15a. It is important to isolate each N-site when sgps contain two or more glycosylated N-sites.

c. Now, column S will have peptides before the 1st glycan; column T will have the 1st glycan and following peptide residues; columns U and V, if applicable, will contain the 2nd and 3rd glycans plus any following peptides.

d. Highlight columns T, U, V. Change "{n}" into "{n}/" to avoid mis-replacement in following step. This is done because {n} is contained in many other bigger glycoforms such as {n{n{h{h{n}}{h{h{n}}}}}}, and can cause unwanted misreplacement of the bigger glycoforms.

   i. Replace "{n}}" with itself, fill cells orange.

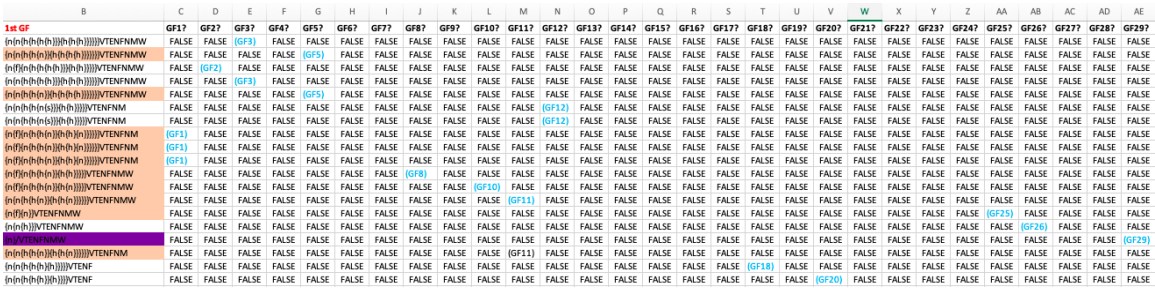
   ii. Find "{n}" in colorless cells and replace with "{n}/". May edit replaced cells with purple fill to better visualize, but color code is optional.

   iii. After performing steps 10a to 10d, columns T, U, and V are fully prepared for the next step. An example is shown in the picture below:

T	U	V
1st GF	**2nd GF**	**3rd GF**
{n{n{h{h{h}}}{h{h{h}}}}}VTENFNMWK		
{n{n}}VTENF		
{n{n{h{h{h}}{h{h}}}}}VTENF		
{n{f}{n{h{h{h}}}{h{h}}}}VTENF		
{n{f}{n{h{h{h}}}{h{h}}}}CTNVN	{n{f}}VTNL	
{n{f}}CTNVN	{n{f}{n{h{h{h}}{h{h}}}}}VTNL	
{n{f}{n{h{h{h}}{h{h}}}}}CTNVN	{n{f}}VTNL	
{n{f}}CTNVN	{n{f}{n{h{h{h}}{h{h}}}}}VTNL	
{n{f}}CTNVNVTNL		
{n{f}{n{h{h{h}}{h{h}}}}}CTNVN	{n}/VTNL	
{n{f}{n{h{h{h{n}}}{h{h{h}}}}}}CTNVNVTNL		
{n{f}}CTNVN	{n{n{h{h{h}}{h{h}}}}}VTNL	
{n}/CTNVN	{n{n}}VTNLKN	{n{n{h{h{n}}{h}}}}ETNTNSSSGGEKM
{n}/CTNVN	{n}/VTNLKNETNTN	{n{n{h{h{n}}{h{n}}}}}SSSGGEKM
{n}/CTNVN	{n}/VTNLKN	{n{n{h{h{n}}{h{n}}}}}ETNTNSSSGGEKM
{n{f}{n{h{h{h}}{h{h}}}}}CTNVNVTNLKNETNTNSSSGGEK		

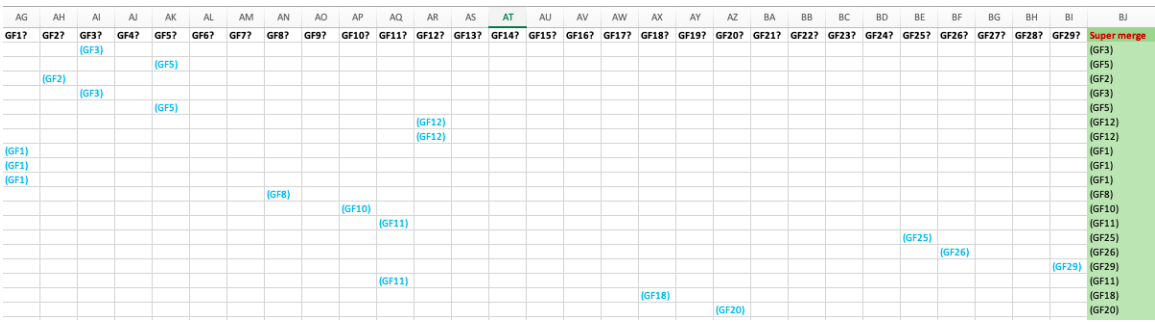
e. Copy column T (1st GF) to column B of sheet named "{} to GFx". This sheet uses formulas to convert glycoforms from "{}" notation to "GF" numbers: eg. C2=IF(ISNUMBER(SEARCH("{n{f}{n{h{h{n}}{h{h}{n}}}}}}",B2)),"

(GF1)"). For example, if {n{f}{n{h{h{n}}{h{h}{n}}}}}} is found in cell B2, cell C2 will display GF1. Otherwise, cell C2 will display the message "FALSE". Check that formulas in columns C-AE reach the end of the list in column B, which could go beyond row 40,000. The picture below shows an example from this sheet:



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
	1st GF	GF1?	GF2?	GF3?	GF4?	GF5?	GF6?	GF7?	GF8?	GF9?	GF10?	GF11?	GF12?	GF13?	GF14?	GF15?	GF16?	GF17?	GF18?	GF19?	GF20?	GF21?	GF22?	GF23?	GF24?	GF25?	GF26?	GF27?	GF28?	GF29?
1	{n{n{h{h{h}}}}{h{h{h}}}}VTENFNMW	FALSE	FALSE	(GF3)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	{n{n{h{h{n}}}{h{h{h}}}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	(GF5)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	{n{f}{n{h{h{h}}}}{h{h}}}VTENFNMW	FALSE	(GF2)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	{n{n{h{h{h}}}}{h{h{h}}}}VTENFNMW	FALSE	FALSE	(GF3)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	{n{n{h{h{n}}}}{h{h{h}}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	(GF5)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	{n{n{h{n{s}}}}{h{h}}}VTENFNM	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF12)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	{n{n{h{n{s}}}}{h{h}}}VTENFNM	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF12)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	{n{f}{n{h{h{n}}}{h{h}{n}}}}VTENFNM	(GF1)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	{n{f}{n{h{h{n}}}{h{h}{n}}}}VTENFNM	(GF1)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	{n{f}{n{h{h{n}}}{h{h}{n}}}}VTENFNM	(GF1)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	{n{f}{n{h{h{n}}}{h{h}{n}}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF8)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	{n{f}{n{h{h{n}}}{h{n}}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF10)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	{n{n{h{n}}}{h{h{n}}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF11)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	{n{f}{n}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF25)	FALSE	FALSE	FALSE	FALSE
15	{n{n{h}}}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF26)	FALSE	FALSE	FALSE
16	{n}VTENFNMW	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF29)
17	{n{n{h{h{n}}}{h{h{h}}}}}VTENFNM	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF11)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	{n{n{h{h{h}}}}{h{h}}}VTENF	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF18)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
19	{n{n{h{h{h}}}{h}}}VTENF	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	(GF20)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

f. Copy columns C-AE (values only) to the replacement area (columns AG-BI). Remove cells containing "FALSE" by replacing "FALSE" with nothing. Autofill "super-merge" formula in column BJ. An example is shown below:



AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ
GF1?	GF2?	GF3?	GF4?	GF5?	GF6?	GF7?	GF8?	GF9?	GF10?	GF11?	GF12?	GF13?	GF14?	GF15?	GF16?	GF17?	GF18?	GF19?	GF20?	GF21?	GF22?	GF23?	GF24?	GF25?	GF26?	GF27?	GF28?	GF29?	Super merge
		(GF3)																											(GF3)
				(GF5)																									(GF5)
	(GF2)																												(GF2)
		(GF3)																											(GF3)
				(GF5)																									(GF5)
											(GF12)																			(GF12)
											(GF12)																			(GF12)
(GF1)																													(GF1)
(GF1)																													(GF1)
(GF1)																													(GF1)
							(GF8)																						(GF8)
									(GF10)																				(GF10)
										(GF11)																			(GF11)
																								(GF25)					(GF25)
																									(GF26)				(GF26)
																												(GF29)	(GF29)
										(GF11)																			(GF11)
																	(GF18)												(GF18)
																			(GF20)										(GF20)

g. The super merge column now contains the list of glycoforms in GFx notation and copy this to column "1st GF" in final table.

h. So far, only the 1st GF (column T from sheet named "posi #, remove dups, pretreat") is converted to the GFx notation. Repeat steps 15e to 15g for 2nd and 3rd GF (column U and V in sheet named "posi #, remove dups, pretreat").

22. The final table should now have all 3 columns of GFs and 3 columns of N-sites filled. Now autofill the "merge 1" list which will combine the cell contents of 1st

site and 1st GF. Repeat for “merge 2” and “merge 3”. The picture below shows an example of this sheet at this stage:

1st GF	2nd GF	3rd GF	1st N site	2nd N site	3rd N site	merge 1	merge 2	merge 3
(GF29)			1	2	3	1(GF29)	2	3
(GF23)			2	3	4	2(GF23)	3	4
(GF27)			3	4	5	3(GF27)	4	5
(GF26)			4	5	6	4(GF26)	5	6
(GF18)			5	6	7	5(GF18)	6	7
(GF18)			6	7	8	6(GF18)	7	8
(GF7)			7	8	9	7(GF7)	8	9
(GF7)			8	9	10	8(GF7)	9	10
(GF14)	(GF23)		9	10	11	9(GF14)	10(GF23)	11
(GF14)	(GF25)		10	11	12	10(GF14)	11(GF25)	12
(GF14)	(GF26)		11	12	13	11(GF14)	12(GF26)	13
(GF14)	(GF28)		12	13	14	12(GF14)	13(GF28)	14
(GF20)	(GF21)		13	14	15	13(GF20)	14(GF21)	15
(GF20)	(GF23)		14	15	16	14(GF20)	15(GF23)	16
(GF20)	(GF25)		15	16	17	15(GF20)	16(GF25)	17
(GF20)	(GF24)		16	17	18	16(GF20)	17(GF24)	18
(GF20)	(GF26)		17	18	19	17(GF20)	18(GF26)	19
(GF20)	(GF28)	(GF21)	18	19	20	18(GF20)	19(GF28)	20(GF21)
(GF5)	(GF25)	(GF29)	19	20	21	19(GF5)	20(GF25)	21(GF29)
(GF5)	(GF27)	(GF28)	20	21	22	20(GF5)	21(GF27)	22(GF28)
(GF5)	(GF28)	(GF27)	21	22	23	21(GF5)	22(GF28)	23(GF27)
(GF5)	(GF29)	(GF25)	22	23	24	22(GF5)	23(GF29)	24(GF25)
(GF9)	(GF23)	(GF29)	23	24	25	23(GF9)	24(GF23)	25(GF29)
(GF9)	(GF27)	(GF26)	24	25	26	24(GF9)	25(GF27)	26(GF26)
(GF9)	(GF26)	(GF27)	25	26	27	25(GF9)	26(GF26)	27(GF27)
(GF9)	(GF29)	(GF23)	26	27	28	26(GF9)	27(GF29)	28(GF23)

23. Once the merge lists are generated, the table on the right will automatically find the number of times each unique combination appears, ie., a matrix of each glycoforms at each N-site. Example shown in picture below:

24. This table needs a final verification step because it assumes that the 2nd Glycan is attached to the 2nd N-site, whereas in reality it might come from the 3rd N-site when the 2nd N-site is not occupied. In this case, correction of glycans from the 2nd N-site to 3rd N site (2-to-3) is necessary. Similarly, if a peptide contains 4 potential N-sites, and only 2 or 3 sites are occupied, it may be required to move 2-to-3, 3-to-4, 23-to-34, or 2-to-4. The verification step is done by checking the peptide sequences between the 1st and 2nd glycosylated N-sites (and between 2nd and 3rd glycosylated N-sites) for "skipped" N-site. Skipped N-site are not occupied by glycans and therefore not interrupted by the spaces in pretreatment (step 15a and 15b).

    a. Go to column X of sheet named "posi #, remove dups, pretreat", which finds the number of glycans in each sgp using the following pre-entered formula: X2=COUNTIF(T2:V2,"*{*"). Autofill column X by double clicking the bottom right corner of cell X2.

    b. Copy columns T and U to columns Y and Z. Remove glycans by replace function. This leaves only the peptide between the 1st and 2nd glycans in column Y and the peptide between the 2nd and 3rd glycans in column Z. Renamed column Y and Z as "peptide after 1st glycan" and "peptide after 2nd glycan". An example of step 13a and 13b is shown in the picture below:

T	U	V	W	X	Y	Z
1st GF	**2nd GF**	**3rd GF**		**# of glycans in sgp**	**peptide after 1st GF**	**peptide after 2nd GF**
{n{n{h{h{h}}}}{h{h{h}}}}VTENFNMWK			prepare for verification	1	VTENFNMWK	
{n{n}}VTENF				1	VTENF	
{n{n{h{h{h}}}{h{h}}}}VTENF			find the # of glycans for each sgp	1	VTENF	
{n{f}{n{h{h{h}}}{h{h}}}}VTENF			formula=COUNTIF(T2:V2,"*{*")	1	VTENF	
{n{f}{n{h{h{h}}}{h{h}}}}CTNVN	{n{f}}VTNL		this formula finds # of cells containing "{"	2	CTNVN	VTNL
{n{f}}CTNVN	{n{f}{n{h{h{h}}}{h{h}}}}VTNL			2	CTNVN	VTNL
{n{f}{n{h{h{h}}}{h{h}}}}CTNVN	{n{f}}VTNL		copy over sgp, 1st GF, 2nd GF	2	CTNVN	VTNL
{n{f}}CTNVN	{n{f}{n{h{h{h}}}{h{h}}}}VTNL		(column S, T, U), values only	2	CTNVN	VTNL
{n{f}}CTNVNVTNL			remove glycans	1	CTNVNVTNL	
{n{f}{n{h{h{h}}}{h{h}}}}CTNVN	{n}VTNL		(replace (*) with nothing)	2	CTNVN	VTNL
{n{f}{n{h{h{n}}}{h{h{h}}}}}CTNVNVTNL			(replace "}", then "/")	1	CTNVNVTNL	
{n{f}}CTNVN	{n{n{h{h{h}}}{h{h}}}}VTNL			2	CTNVN	VTNL
{n}/CTNVN	{n{n}}VTNLKN	{n{n{h{h{n}}}{h{n}}}}ETNTNSSSGGEKM	rename sgp as "**pep before 1st GF**"	3	CTNVN	VTNLKN
{n}/CTNVN	{n}/VTNLKNETNTN	{n{n{h{h{n}}}{h{n}}}}SSSGGEKM	rename 1st GF as "**pep after 1st GF**"	3	CTNVN	VTNLKNETNTN
{n}/CTNVN	{n}/VTNLKN	{n{n{h{h{n}}}{h{n}}}}ETNTNSSSGGEKM	rename 2nd GF as "**pep after 2nd GF**"	3	CTNVN	VTNLKN
{n{f}{n{h{h{h}}}{h{h}}}}CTNVNVTNLKNETNTNSSSGGEK			copy columns X, Y, Z, AA to verification sheet	1	CTNVNVTNLKNETNTNSSSGGEK	

    c. Copy columns X, Y, Z to columns B, C, D in sheet named "verification".

25. Copy columns B, C, D to columns H, I, J. Highlight all three columns of H, I, J and go to DATA>remove duplicates, creating a workable “short list” of the original columns B, C, D. Ignore rows in columns H, I, J if “# of glycans” (cells in column H) is 1 because only glycopeptides with two or more glycans are potentially affected by the issue of inaccurate glycan-site matching.

26. Manually check the remaining unique peptides in columns I and J for any “skipped” N sites. A peptide is considered “skipped” if it contains two or more N-sites within its sequence, as “non-skipped” peptides should only contain one N-site. Highlight the identified peptides with 1 skipped N-site in red and the peptides

I	J
pep after 1st GF	**pep after 2nd GF**
CTNVN	VTNL
CTNVN	VTNLK
CTNVNVTNLKNETNTN	SSSGGEK
CTNVNVTNLKN	ETNTNSSSGGEK
CTNVN	VTNLKNETNTNSSSGGEK
CTNVN	VTNLKNETNTNSSSGGEKM
VTNLKNETNTN	SSSGGEK
VTNLKN	ETNTNSSSGGEK
VTNLKNETNTN	SSSGGEKM
VTNLKN	ETNTNSSSGGEKM
ETNTN	SSSGGEK
CSFN	VTTLIRNK
TSYTLINCN	SSTITQACPK
GSLAEEDIVIRSEN	F
GSLAEEDIVIRSEN	FTDNAK
VSIEINCTRPNN	NTRK
VSIEIN	CTRPNNNTRK
CTRPNN	NTRK
STQLFN	STW
GTWKNTEGADNN	ITLPCRIK
ASWSN	RSQDY
RSQDYIWNN	M
CTNVNVTNLKN	ETNTN
CTNVN	VTNLKNETNTN
CTNVN	VTNLKN
VTNLKN	ETNTN
VSIEIN	CTRPNN
skip 1 site (2nd N out of 3 or 4 Ns skipped) "2-3-skip-"	skip 1 site (3rd N out of 4 Ns skipped) "3-4-skip-"
skip 2 sites (2nd and 3rd N out of 4 Ns skipped) "2-4-skip-"	

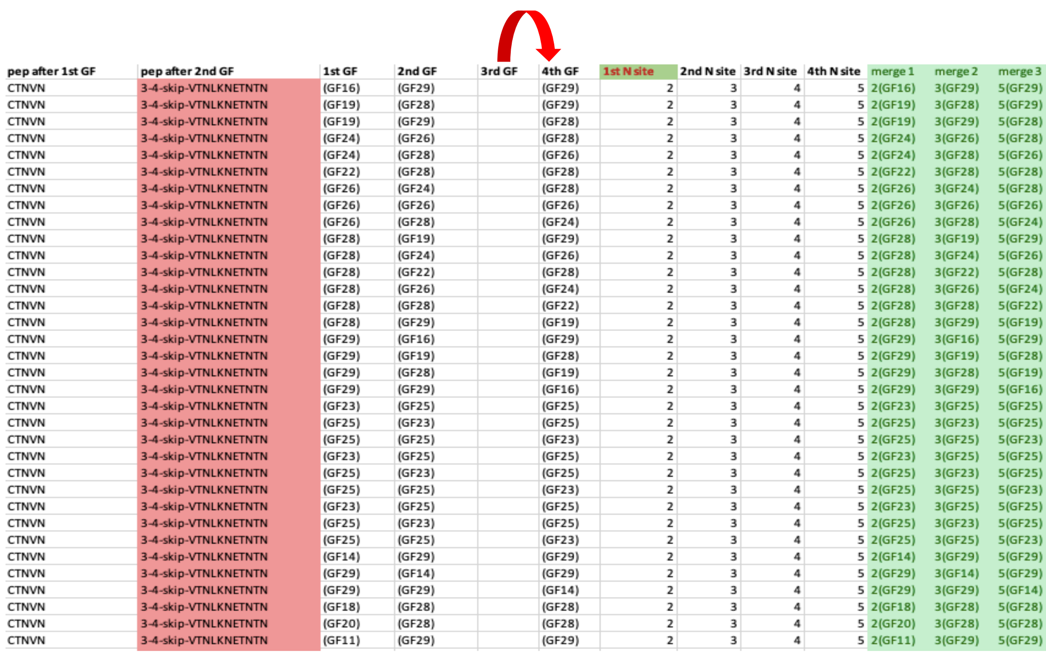
with 2 skipped sites in purple. Example is shown below using B4:

27. Go back to original columns C and D, and use replace function to add “2-3-skip-”, “2-4-skip-”, or “3-4-skip-” in front of any peptides containing skipped N sites.

28. Copy modified columns C and D to columns B and C of sheet named “final table”.

    a. Select columns B-K (all columns before the merged lists), and then sort by column B. This brings the “2-3-skip” and “2-4-skip”-tagged rows to the top. Manually transfer the cell contents from column “2nd GF” and “3rd GF” to “3rd GF” and “4th GF”, respectively, if the row is “2-3-skip”-tagged. Transfer cell contents from column “2nd GF” to “4th GF” if the row is “2-4-skip”-tagged.

    b. Select columns B-K (all columns before the merged lists), and then sort again by column C. This brings the “3-4-skip”-tagged rows to the top. Transfer cell contents from column “3rd GF” to “4th GF” if the row is “3-4-skip”-tagged. When this step is completed, the lists should resemble the example shown in the picture below:



pep after 1st GF	pep after 2nd GF	1st GF	2nd GF	3rd GF	4th GF	1st N site	2nd N site	3rd N site	4th N site	merge 1	merge 2	merge 3
CTNVN	3-4-skip-VTNLKNETNTN	(GF16)	(GF29)		(GF29)	2	3	4	5	2(GF16)	3(GF29)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF19)	(GF28)		(GF29)	2	3	4	5	2(GF19)	3(GF28)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF19)	(GF29)		(GF28)	2	3	4	5	2(GF19)	3(GF29)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF24)	(GF26)		(GF28)	2	3	4	5	2(GF24)	3(GF26)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF24)	(GF28)		(GF26)	2	3	4	5	2(GF24)	3(GF28)	5(GF26)
CTNVN	3-4-skip-VTNLKNETNTN	(GF22)	(GF28)		(GF28)	2	3	4	5	2(GF22)	3(GF28)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF26)	(GF24)		(GF28)	2	3	4	5	2(GF26)	3(GF24)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF26)	(GF26)		(GF26)	2	3	4	5	2(GF26)	3(GF26)	5(GF26)
CTNVN	3-4-skip-VTNLKNETNTN	(GF26)	(GF28)		(GF24)	2	3	4	5	2(GF26)	3(GF28)	5(GF24)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF19)		(GF29)	2	3	4	5	2(GF28)	3(GF19)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF24)		(GF26)	2	3	4	5	2(GF28)	3(GF24)	5(GF26)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF22)		(GF28)	2	3	4	5	2(GF28)	3(GF22)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF26)		(GF24)	2	3	4	5	2(GF28)	3(GF26)	5(GF24)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF28)		(GF22)	2	3	4	5	2(GF28)	3(GF28)	5(GF22)
CTNVN	3-4-skip-VTNLKNETNTN	(GF28)	(GF29)		(GF19)	2	3	4	5	2(GF28)	3(GF29)	5(GF19)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF16)		(GF29)	2	3	4	5	2(GF29)	3(GF16)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF19)		(GF28)	2	3	4	5	2(GF29)	3(GF19)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF28)		(GF19)	2	3	4	5	2(GF29)	3(GF28)	5(GF19)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF29)		(GF16)	2	3	4	5	2(GF29)	3(GF29)	5(GF16)
CTNVN	3-4-skip-VTNLKNETNTN	(GF23)	(GF25)		(GF25)	2	3	4	5	2(GF23)	3(GF25)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF23)		(GF25)	2	3	4	5	2(GF25)	3(GF23)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF25)		(GF23)	2	3	4	5	2(GF25)	3(GF25)	5(GF23)
CTNVN	3-4-skip-VTNLKNETNTN	(GF23)	(GF25)		(GF25)	2	3	4	5	2(GF23)	3(GF25)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF23)		(GF25)	2	3	4	5	2(GF25)	3(GF23)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF25)		(GF23)	2	3	4	5	2(GF25)	3(GF25)	5(GF23)
CTNVN	3-4-skip-VTNLKNETNTN	(GF23)	(GF25)		(GF25)	2	3	4	5	2(GF23)	3(GF25)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF23)		(GF25)	2	3	4	5	2(GF25)	3(GF23)	5(GF25)
CTNVN	3-4-skip-VTNLKNETNTN	(GF25)	(GF25)		(GF23)	2	3	4	5	2(GF25)	3(GF25)	5(GF23)
CTNVN	3-4-skip-VTNLKNETNTN	(GF14)	(GF29)		(GF29)	2	3	4	5	2(GF14)	3(GF29)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF14)		(GF29)	2	3	4	5	2(GF29)	3(GF14)	5(GF29)
CTNVN	3-4-skip-VTNLKNETNTN	(GF29)	(GF29)		(GF14)	2	3	4	5	2(GF29)	3(GF29)	5(GF14)
CTNVN	3-4-skip-VTNLKNETNTN	(GF18)	(GF28)		(GF28)	2	3	4	5	2(GF18)	3(GF28)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF20)	(GF28)		(GF28)	2	3	4	5	2(GF20)	3(GF28)	5(GF28)
CTNVN	3-4-skip-VTNLKNETNTN	(GF11)	(GF29)		(GF29)	2	3	4	5	2(GF11)	3(GF29)	5(GF29)

29. The final table shown in step 12 should be automatically refreshed. This table will accurately reflect the number of times that each glycoform-N-site combination appears.

30. Often, some N-sites are too close to be separated by enzyme digestion and remain on the same peptide fragment. Mass spectrometry cannot distinguish the sites from which glycans are cleaved. GlycoPAT outputs all possibilities, which double counts or triple counts the actual number of glycans present for these sites. These sites should be grouped into “clusters” simply by removing the duplicated information before graphing or statistical analysis.

31. This is the end of this protocol. All further actions should be based on the final table generated in step 19.

Curriculum Vitae

Name: Yingxue Sun

Post-secondary Education and Degrees:

The University of Western Ontario

London, Ontario, Canada

2014-2018 B.Sc. Hons. Double-major in genetics and cell biology

The University of Western Ontario

London, Ontario, Canada

2018- M.Sc. candidate in microbiology and immunology

Honours and Awards:

Dean's Honor List

2016-2017

The Western Scholarship of Excellence

2014-2015

Related Work Experience

Summer student

McMaster University

2012-2014

Publications:

Yang, G., Geng, X.R., Liu, Z.Q., Liu, J.Q., Liu, X.Y., Xu, L.Z., Zhang, H.P., **Sun, Y.X**., Liu, Z.G. and Yang, P.C. (2015). Thrombospondin-1 (TSP1)-producing B cells restore antigen (Ag)-specific immune tolerance in an allergic environment. Journal of Biological Chemistry, 290(20), 12858-12867.