bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Manuscript 2 of 3

Sialoglycan microarray encoding reveals differential sialoglycan binding of phylogenetically-related bacterial AB_5 toxin B subunits

Aniruddha Sasmal[1,2], Naazneen Khan[1,2], Zahra Khedri[1,2], Benjamin P. Kellman[3], Saurabh Srivastava[1,2], Andrea Verhagen[1,2], Hai Yu[5], Anders Bech Bruntse[3], Sandra Diaz[1,2], Nissi Varki[1,2], Travis Beddoe[6], Adrienne W. Paton[7], James C. Paton[7], Xi Chen[5], Nathan E. Lewis[1,3] and Ajit Varki[1,2,4]

[1]*Glycobiology Research and Training Center,* [2]*Department of Cellular & Molecular Medicine,* [3]*Department of Pediatrics, and* [4]*Department of Medicine, University of California, San Diego, La Jolla, CA-92093,* [5]*Department of Chemistry, University of California, Davis, CA, 95616, USA,* [6]*Department of Animal, Plant and Soil Science and Centre for AgriBioscience, La Trobe University, Bundoora, Victoria 3086, Australia, and* [7]*Research Centre for Infectious Diseases, Department of Molecular and Biomedical Science, University of Adelaide, Adelaide, SA 5005, Australia.*

Running Title: Sialoglycan microarray encoding of AB_5 toxin B subunits.

*Address correspondence to:

Ajit Varki
Email: a1varki@ucsd.edu
UCSD School of Medicine
La Jolla, CA 92093-0687

Keywords

Sialic acid, glycobiology, *N*-acetylneuraminic acid, *N*-glycolylneuraminic acid, glycan microarray, glycan codes, *Yersinia enterocolitica*, bacterial AB_5 toxin, B subunit, glycan binding.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract (250/250 words)

Vertebrate sialic acids (Sias) display much diversity in modifications, linkages and underlying glycans. Slide microarrays allow high-throughput analysis of sialoglycan-protein interactions. The preceding paper used ~150 structurally-defined sialyltrisaccharides with various Sias and modified forms at non-reducing ends, to compare pentameric sialoglycan-recognizing bacterial toxin B subunits. Unlike the poor correlation between B subunits and species phylogeny, there is stronger correlation with Sia types prominently expressed in susceptible species. Further supporting this pattern we report a B subunit(YenB) from *Yersinia enterocolitica* (broad host range) recognizing almost all sialoglycans in the microarray, including 4-*O*-acetylated-Sias not recognized by a *Y.pestis* orthologue(YpeB). Differential Sia-binding patterns were also observed with phylogenetically-related B subunits from *Escherichia coli*(SubB), *Salmonella* Typhi(PltB), *S.* Typhimurium(ArtB), extra-intestinal *E.coli*(*Ec*PltB), *Vibrio cholera*(CtxB), and cholera family homologue of *E. coli*(EcxB). Given library size, data sorting and analysis posed a challenge. We devised a 9-digit code for trisaccharides with terminal Sias and underlying two monosaccharides assigned from the non-reducing end, with three digits assigning a monosaccharide, its modifications, and linkage. This code allows logical sorting, motif searching of results, and optimizes printing. While we developed the system for the >113,000 possible linear sialyltrisaccharides, we note that a biantennary *N*-glycan with two terminal sialoglycan trisaccharides could have $>10^{10}$ potential combinations and a triantennary *N*-glycan with three terminal sequences, $>10^{15}$ potential combinations. While all possibilities likely do not exist in nature, sialoglycans encode enormous diversity. Thus, while glycomic approaches address these challenges, naturally-occurring toxin B subunits are simpler tools to track the dynamic sialome in biological systems.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Glycans play critical roles in various structural and modulatory functions in cells. Cell surface glycans are particularly important as they are involved in various biological processes, like cell signaling, host-pathogen interaction, and immune interactions (1). Sialic acids (Sias), nine-carbon backbone monosaccharides, are the most prominent terminal monosaccharides found in mammalian cell surface glycomes. The sialome is defined as the ''*total complement of sialic acid types and linkages and their mode of presentation on a particular organelle, cell, tissue, organ or organism—as found at a particular time and under specific conditions*''(2). The huge diversity in sialomes arises from the monosaccharide identity, functional modification, and linkage of Sia. Besides the two most common mammalian Sias, *N*-acetylneuraminic acid (Neu5Ac) and *N*-glycoylneuraminic acid (Neu5Gc), the Sia family includes ketodeoxynononic acid (Kdn), and neuraminic acid (Neu), which are rarely found in mammalian glycomes. Glycans are one of the major classes of macromolecules in nature. Due to the structural complexity of glycans compared to the simpler linear sequences of nucleotides and peptides, the development and availability of glycan informatic resources, such as user-friendly nomenclature, structure motif finder, and database are still limited, hindering the advancement of glycan research (3, 4). Unlike nucleic acids and proteins, glycan sequences are also comprised of a large number of possible monosaccharide residues, variable linkages, and branching possibilities (5, 6). The possibility of multiple functional modification and linkages on seven out of the nine carbons in Sias made it particularly challenging to establish a functional yet convenient representation of sialosides. The diversity of sialosides increases further when the underlying glycan structures are taken into account.

To partly address this issue, an abbreviated text-based nomenclature describing the glycan structure was established by the IUPAC-IUBMB (International Union for Pure and Applied Chemistry and International Union for Biochemistry and Molecular Biology) (7). However, this three-letter coding system is ambiguous and not easily transferable to a machine-readable format. Later LINUCS (Linear Notation for Unique description of Carbohydrate Sequences) was developed to address the IUPAC limitations (8), but it has not had universal acceptance and applications. Several efforts have since been made to develop a comprehensive representation of complex glycan structures (9–12) and their associated reaction rules (13), and establish an algorithm for the mining of high-complexity glycan motifs involved in biological processes (14). Early attempts have been made to implement useful identifiers for explicitly describing the glycan structures in databases like Complex Carbohydrate Structure Database, CarbBank, CFG etc. (13, 15–19). Recently, assignment of unique glycan structure identifiers in the GlyTouCan repository (20) have been increasingly accepted by the glycomics community. A graphical representation of glycan structures by the 'Symbol Nomenclature for Glycans' (SNFG) system has also been standarized (21).

Automated and high-throughput analytical techniques, such as glycan microarray produces a large amount of glycan binding data (22, 23). Efforts, albeit limited, have been made to establish proper nomenclature and motif finders for in-depth study of structure and function relationships of glycans (24–26). For detailed interpretation of vast microarray datasets, we establish here a suitable structural representation of glycans in a machine-readable digital format.

Historically, our groups have been focusing on identifying novel biological roles of terminal Sias and consequently, and over the years we developed a sialoglycan-focused microarray platform to explore the novel biomolecular interactions with the sialosides (27–29). Our extensive sialoglycan inventory includes mostly sialyltrisaccharides with various modification and linkages of Sia. The ever increasing sialoglycan microarray library demands an efficient coding-based structure representation system to streamline the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

printing, sorting, analysis, and motif-finding processes. A coding-based systematic representation of complex *N*-linked glycans has been reported, however the functional modification on the terminal Sias was not addressed in detail (30). Hence, we set out to develop a simple yet versatile data format for the representation of mainly sialyltrisaccharide structures that we use for our elaborate sialoglycan microarray library and hope this will benefit the glyco-community as a whole.

The characterisation of an AB_5 toxin B-subunit from *Y. pestis* in the preceding paper (accompanying manuscript Khan *et al.*) also made us inquisitive about the Sia preferences of other bacterial AB_5 toxin B subunits, which are responsible for cell surface receptor recognition. A systematic screening of phylogeny-based bacterial toxin B subunits on the sialoglycan microarray was pursued to identify the diversity in sialoglycan-binding pattern.

Finally, the newly developed coding system can be utilized to calculate the number of naturally occurring sialyltrisaccharide structures in mammalian sialomes. Although the complexity and the diversity of the sialome make the calculation of the total number of sialoside structures impossible, we belive that our coding approach will help to address the long-standing question regarding the number of possible glycan structures.

Results and Discussion

A coding system for individual sialoglycans on the microarray. The importance and coverage of high-throughput glycan microarray technology depend on the number and diversity of glycans present in the specific library. Over the years we developed a library of sialosides for microarrays. A series of natural and non-natural sialoglycans (see SI for full list) were synthesized using chemoenzymatic methods (31). We restricted the sialoside structure primarily in the form of defined trisaccharides, as addressing all of the diverse structures of sialosides in nature is beyond anyone's scope. Except for a few non-sialosides, the terminal monosaccharides used are Neu5Ac, Neu5Gc, and Kdn along with their natural modifications. The terminal Sias at the non-reducing end are linked with the underlying glycan structures in α2-3, α2-6, and α2-8 fashion. Terminal Sias are linked to the underlying glycan structures that are usually known to occur in nature (precisely in mammals), for example, Type I (Galβ1-3GlcNAcβ), Type II (Galβ1-4GlcNAcβ), Type III or T-antigen (Galβ1-3GalNAcα), Type IV (Galβ1-3GalNAcβ), Type VI (Galβ1-4Glcβ) (32), Lewis antigens, Tn-antigen, sulfated GalNAc, and ganglioside-type structures. Initially, the individual sialoglycan was assigned with a number (Glycan ID) in the order they were synthesized and stored.

The assigned glycan IDs listed in the table (SI Table *S1*) are semi-random and do not help in streamlining the sorting and analysis of the microarray data. To address these issues, we developed a new coding system. The purpose of this system was to easily order the glycans in various logical ways during spreadsheet analyses and identifying glycan-binding motifs. Ten-digit machine-readable codes for individual glycans were devised (Table 1). Due to its terminal position, most of the biological interest is in the identity of the Sia. Thus, the coding of the glycans starts from the non-reducing terminal Sias followed by their modifications and linkages. Subsequently, the underlying two monosaccharides (or more) were assigned from the non-reducing end, each with a digit indicating the monosaccharide type, its modifications, and linkage.

The first digit in the 10-digit code represents the terminal Sia identity. Being the most prevalent Sia in mammalian glycomes, Neu5Ac was assigned number *1* followed by its two most important *O*-acetyl modifications at C-4 (number *2*) and C-9 (number *3*) positions. It is noteworthy that Neu5Ac with only C7-*O*-acetyl modification is not stable (33) and hence it was not considered as a primary Sia form. Like Neu5Ac,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Neu5Gc and its *O*-acetyl-modified versions were then assigned the number *4* to *6*. The less abundant Kdn and its *O*-acetyl modifications were assigned afterwards to occupy the numbers *7* to *9*. To accommodate the non-sialosides in our library, we decided to assign number *0* for the 'no sialic acid' at the non-reducing end of the glycan. We extended the assignment system to accommodate less abundant (or not present in our library yet). An alphabet (small letter) was assigned to each of the neuraminic acid (Neu), legionaminic acid (5,7-di-*N*-acetyl legionaminic acid, Leg5,7Ac_2), and other nonulosonic acids mostly present in prokaryotes. It is noteworthy that while the existence of 'naked' Leg is theroretically possible, all the naturally occurring legionaminic acids reported thus far are *N*-acylated at C-5 and C-7 positions. These sialoglycans were assigned with letters starting from *a* and filling downwards as the library expands. The second digit in the coding system was to identify the modification on the Sia. We assigned numbers *1* through *9* for each type of naturally occurring *O*-acetyl, lactone, *O*-lactyl, 5-OMe, and 8-OMe modifications on the Sia. To the best of our ability, we tried to rationalize the assignment of the number by matching the carbon position of the specific Sia modification. Number *9* was assigned to the (7/8),9-di-*O*-acetyl modification where the position for one of the *O*-acetyl groups is not confirmed due to the *O*-acetyl migration among C-7, C-8, and C-9 positions (33). The 1,5-lactam, a modification generally found in neuraminic acid (Neu) was assigned with alphabet *a*. Next in the column, the third digit represented the various types of natural linkages connecting the C-2 position of Sia to the next underlying monosaccharide unit in alpha stereochemical fashion. To date, no beta linkage was reported to occur at the C-2 position of the Sia (except in the CMP-monosaccharide state). Rationally, the α2-3 connection was assigned to number *3*, α2-6 connection to number *6* and so on. To make our numbering system more self-explanatory, we did not assign number *1* and *2* to any of the linkages, the obvious reason being the absence of α2-1 and α2-2 Sia connections. In total the first three digits in the coding system were sufficient to properly express most Sias at the non-reducing end. Similarly, the next three digits enumerated the identity, modification/branching, and the linkage of the next underlying monosaccharide. In a poly/oligosialic acid, the Sia is linked to another Sia in α2-8-fashion and for that reason we included Sia in the list of second sugar unit (fourth digit). To keep the numbering system in sync, the Neu5Ac, Neu5Gc, and Kdn in this column were assigned to the same numbers/alphabets as in the first digit place (second column of the Table 1). We assigned leftover numbers to the naturally occurring monosaccharides such as galactose (Gal), *N*-acetylgalactosamine (GalNAc), and *N*-acetylglucosamine (GlcNAc). Apart from the listed ones, the occurrence of any other monosaccharides immediate to the terminal Sia is so far unknown. We did not use the numbers 3, 5, and 9 for the fourth digit place as there are no known linkages of non-terminal sialic acid through C-3, C-5, or C-9 positions (except on some bacterial capsular polysaccharide). Next, the branching at the second sugar unit was represented by the fifth digit. The branching monosaccharides were assigned along with their linkages. The Fucα1-2 (found in $Lewis^Y$, $Lewis^B$, and human blood group antigens); Galα1-3 (found in B-blood group antigen/alpha-Gal), GalNAcβ1-4 (found in ganglio-series gangliosides), and 3/6-*O*-sulfation were assigned with the numbers that were in sync with the number assignments in other columns. Although we focused on trisaccharides, we took the liberty to accommodate some important natural glycans with longer side chains such as GM1 and Fuc-GM1 structures. The Galβ1-3GalNAcβ1-4 and Fucα1-2Galβ1-3GalNAcβ1-4 branches were assigned with letter *e* and *h* respectively. Next, the sixth digit was assigned to the linkage that connects the reducing end of the second sugar and the non-reducing end of the third sugar unit. The number assignment was in accordance with the previous linkage assignments listed in the fourth column (the third digit). The identity of the third sugar in the linear chain was denoted with the seventh digit. The number assignment of the monosaccharide for the seventh digit place remained the same as in the second sugar unit (fourth digit). The only addition was the glucose (Glc) monosaccharide, which was assigned with number *5*. The branching on the third sugar was designated in the eighth digit place in the coding system. The

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

branched Sias on the third sugar were assigned with the numbers similar to those described for the first and the fourth digits. Again, the naturally occurring glycans and sulfated modifications were assigned with the numbers. In the eighth digit place, a series of di- and tri-saccharide branching (as found in Core 2, GD1) was considered and each was assigned with a letter. The linkage at the reducing end of the third sugar was designated in the ninth digit place. It is important to remember that the linkages had to be between two sugar units. Generally, the third sugar unit in our library was attached with an aliphatic linker at the reducing end and hence the ninth digit was always *0* (none) in our case. We assigned an additional tenth digit place to describe the linker connection or any further monosaccharides that are not present at this moment in our library. The α-linked linker was assigned with number *8* and β-linked linker is assigned with number *9*.

Utilizing the coding system, the sialoglycans in our library can be conveniently assigned with a 10-digit code. Fig. 1 demonstrates the representation of sialoglycan in a machine-readable numerical format. See supporting information (Table S1) for the full list of the numerically encoded sialoglycans.

Application of the coding system for logical sorting of array results. One of the major reasons behind developing the coding system was to make the microarray data sorting easy and logical. Fig. 2 shows the binding of polyclonal chicken anti-Neu5Gc antibody on the sialoglycan microarray experiment. When the data were plotted against the semi-random glycan ID (Fig. 2A), the interpretation of data became clumsy and lacked a logical explanation. Once we applied the newly developed coding system on the glycans and sorted the data based on the 10-digit code, the chart became very clear and self-explanatory. The 10-digit codes were quickly arranged in an ascending order of the new glycan ID (in Microsoft Excel) to logically group the Neu5Gc-glycans (and its modified versions) together. The data plotted afterwards (as bar diagram) gave a clear visualization of the specificity of the antibody towards the Neu5Gc-epitopes (Fig. 2B).

In another example, we demonstrated that the 10-digit code not only can be helpful in arranging the data in ascending or descending order, but the numbers at the individual digit place can also be grouped together to convey a meaningful result. To demonstrate the technique, we tested the biotinylated Sambucus Nigra lectin (SNA) on the sialoglycan microarray. The binding dataset was organized by grouping the linkages of the terminal Sia unit, i.e., the third digit. From the plot (Fig. 3), it was clearly demonstrated that SNA prefers to bind α2-6-linked Sia epitopes.

Application of the coding system for motif searching of results. Among the versatile applications of the coding system, one important application was to search for the glycan motif responsible for binding to the lectins, proteins, and antibodies. Although our focus was on the terminal Sia, the underlying glycan structure plays a critical role that affects the binding pattern of the glycan-binding biomolecules. It was well-known that *Maackia amurensis* Lectin II (MAL II) recognizes the terminal Sia linked to the underlying glycan structures in α2-3 linkage. The motif finding based on the new coding system also helped us to confirm the dependence of the binding specificity on the underlying glycan structures. We tested biotinylated MAL II on our sialoglycan microarray and a heat-map is plotted (Fig. 4). The top 20 strongest signals (average relative fluorescence unit, Ave RFU) obtained from the MAL II-binding on the microarray were listed along with the numerical code of the sialoglycan epitopes. It became obvious that all the strongly-binding epitopes bore the same number sequence from third to seventh digit places, which eventually indicated that the Sia α2-3-linked to the Galβ3GalNAcα (sialyl T-antigen on O-glycan) epitope was the key motif required for MAL II binding. The coded number made the motif recognition much easier compared to finding the motif by reading the structures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discovery of phylogenetically related bacterial AB_5 toxin B subunit with broader Sia-preference. In the companion publication (accompanying manuscript Khan *et al.*), we reported a broad range of Sia-binding interaction of an AB_5 toxin B subunit homologue of *Yersinia pestis* (YpeB) on slide microarray. However, our extensive Sia repertoire revealed that YpeB binds to various modified Neu5Ac and Neu5Gc-terminated glycans, except when the terminal Sia was modified with an *O*-acetyl group at the C-4 position or *O*-methylated at the C-5-*N*-glycolyl or C-8 position. Consequently, the idea of using YpeB as a universal Sia-binding probe had to be dismissed. These shortcomings led us to screen the glycan-binding B subunits of AB_5 toxins from phylogenetically-related bacteria in search of a broader Sia binding probe.

Yersinia genus encompasses a heterogenous collection of Gram-negative facultative non-sporulating anaerobic bacteria that belongs to the Enterobacteriaceae family. *Yersinia* species are divided into different serological groups based on the antibody reactions to different lipopolysaccharides (34). Only three species of *Yersinia*, *Y. enterocolitica*, *Y. pestis*, and *Y. pseudotuberculosis*, have been known to infect humans. Enteropathogenic *Y. enterocolitica* is a food-borne pathogen, which is the most prevalent in humans, causing a broad range of gastrointestinal syndromes and may cause septicemia and meningitis (35–37). Like *Y. pestis*, *Y. enterocolitica* also has a very broad range of host compatibility. Besides humans, the bacteria have also been isolated from rodents, domestic animals (e.g., cattle, pigs, dogs, sheep), and other animals like horse and deer (38). However, the severity of induced gut-associated diseases differs significantly between hosts.

YenB from *Yersinia enterocolitica* recognizes all sialoglycans in the microarray. Pathogen-glycan interactions are important both for colonization of host surfaces and recognition and entry of cells by toxins. Earlier studies reported that plasmid-bearing *Y. enterocolitica* adhere to purified small intestinal mucins from rabbits and humans (39, 40). The binding of the virulent bacteria with mucin was inhibited by purified mucin oligosaccharides and thus suggests the involvement of the glycan moiety in the cell specific interaction. Given the phylogenetic proximity of *Y. pestis* and *Y. enterocolitica* (Fig. 5) and the reported binding affinity of *Y. enterocolitica* with mucin carbohydrates, we anticipated that the *Y. enterocolitica* AB_5 toxin B subunit homologue YenB should recognize diverse Sia epitopes.

YenB (expressed with a C-terminal His6 tag) was purified from recombinant *E. coli* BL21, by Ni^{2+}-NTA chromatography. The purified YenB protein was then tested for Sia binding using the sialoglycan microarray platform. The sialoglycan coding system was used to rationally sort and plot the data for convenient interpretation. The microarray result (Fig. 6) demonstrated a broader Sia recognition by YenB compared to its phylogenetically-related homologue YpeB from *Y. pestis*. Interestingly, YenB showed a strong binding response against almost all types of Sia in our inventory including the 4-OAc, 5-OMe, and 8-OMe Sia, which YpeB failed to recognize. YenB bound to the Sia irrespective of the linkages (α2-3; α2-6; and α2-8), modifications (4-OAc, 9-OAc, and 8-OMe), and underlying glycan structures. Although YenB recognized the Neu5Ac and Neu5Gc-glycans promiscuously, it did not recognize Kdn-glycans. The lack of binding towards the non-sialosides established that the binding of YenB is not promiscuous towards any glycan epitope.

Structure modeling study of YenB and development of YenB mutant for testing in the microarray: Since the crystal structure of YenB is not yet known, we constructued a structural model of YenB using the crystal structure of SubB (3DWP) as a template. Structural alignment of SubB with YenB (Fig. 7A), and YenB with YpeB (Fig. 7B) showed that YenB adopts the conserved OB fold which is typical of the B subunits from the AB_5 toxin family.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparing the Neu5Gc-binding site of SubB with the proposed glycan-binding site within YenB reveals two major differences (Fig. 8A). Firstly, some of critical SubB residues required for binding are not conserved suggesting that glycan binding is different. Secondly the glycan-binding site in YenB is located below the position of glycan-binding site SubB as binding residues in YenB are shifted down. Similar comparison to the YpeB glycan-binding site (Fig. 8B) also revealed a significant departure from the conservation of the amino acid residues at the binding site of YenB. Notably, the YenB has a tyrosine (Y8) residue near the C-4 position of the sialic acid whereas the YpeB and SubB binding pockets have methionine residue (M30 and M10 respectively). This could be a possible reason for the differential 4-OAc sialoglycan binding prefence between YenB versus YpeB and SubB.

Since the crystal structure of YenB is not yet known, we predicted the active sites of this toxin with the available protein structure and sequence of SubB. It is known that serine residues are critical for binding to Neu5Ac and tyrosine interacts with the extra hydroxyl group of Neu5Gc and are thus critical for binding to sialic acids (41). Upon aligning the sequences of *E. coli* SubB and *Y. enterocolitica* YenB, we predicted that the mutation of conserved serine residue (S10) might adversely affect the glycan binding preferences. Additionally, YenB binding site region has isoleucine (I82) as opposed to the conserved tyrosine (Y78) in the SubB (Fig. 8A). Notably, for YenB the residue adjacent to the I82 is a tyrosine (Y81), which is conserved in both species. To ensure a complete disruption of sialoglycan binding by the YenB, we mutated serine to alanine (S10A) and tyrosine to phenylalanine (Y81F). The recombinant YenB-mutant was expressed and the protein was purified according to the similar method described for the YenB. The His-tagged YenB-mutant protein was tested on the microarray and no binding with the sialosides or the non-sialosides was observed whatsoever (SI Fig. *S1*).

Phylogenetic comparison of YenB with other AB_5 toxin B subunits (YpeB, SubB, PltB, ArtB, *Ec*PltB, CtxB, and EcxB). The broad Sia binding propensity of YenB motivated us to explore other phylogenetically-related bacterial AB_5 toxin B subunits. In the accompanying paper (accompanying manuscript, Khan *et al.*), we briefly described the Sia-binding pattern of the Shiga toxogenic *E. coli* Subtilase cytotoxin B subunit (SubB), which shares 59% identity with YpeB. In addition, Sia-binding preferences of B subunits from closely related pathogenic *Salmonella enterica* serovars, *S.* Typhi (PltB) and *S.* Typhimurium (ArtB) were also tested on the sialoglycan microarray platform. Interestingly, a poor correlation was found between the phylogeny of known B subunit sequences in available genomes and bacterial species phylogeny (based on 16sRNA). For example, based on B subunit phylogenetic tree inferences, YenB shares its clade with ArtB (54.24% sequence identity) but not with YpeB (Fig. 5).

Other than the aforementioned B subunits, we found three other closely related B subunits (Fig. 5), which might show interesting Sia-binding patterns. Some pathogenic extra-intestinal *E. coli* strains produce a pertussis-like AB_5 toxin (*Ec*PltAB) (42), related to the toxins from typhoidal and nontyphoidal *Salmonella* serovars. *Ec*PltAB toxin shares 70-80% sequence identity with *S.* Typhi PltAB. The B subunit of the toxin (*Ec*PltB) also shares 69% sequence identity with ArtB from *S.* Typhimurium. A glycan array binding study of *Ec*PltB displayed avidity for a broad range of branched eukaryotic glycans. However, the effect of modifications on the terminal Sia in *Ec*PltB binding was less explored (42).

Cholera is a clinical-epidemiologic syndrome caused by *Vibrio cholerae* that induces severe and at times fatal diarrhea (43), largely attributable to production of cholera toxin (CtxAB), the B subunit of which (CtxB) binds to gangliosides like GM1 (44). It has been reported that the Sia termini on the gangliosides play an important role in target cell recognition (45), although there is also evidence for alternate fucose recognition (46, 47) (which we did not study here). The cholera-like toxins are the largest subgroup of the B_5

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

toxin family. A novel cholera-like toxin was identified from a clinical isolate of *E. coli* (EcxAB) derived from diarrheal patients (48). Unlike other cholera-related AB_5 toxins, the unique mode of pairing of A -subunit to the B subunit categorizes the EcxAB toxin in a novel AB_5 hybrid cholera-like toxin family. Although the A subunits EcxA and CtxA are structurally and functionally very distinct, the B subunits EcxB and CtxB share 63% sequence identity and almost complete structural homology (49). Like CtxB, EcxB was also reported to show strong binding preference towards Neu5Ac-associated ganglioside GM1 and fucosyl-GM1 structures (49). It was therefore of interest to determine whether the binding pattern changes upon substitution of Neu5Ac with Neu5Gc, a naturally less abundant variant which is not present in human (50).

Comparison of sialoglycan binding pattern of phylogenically-related B-subunits (YenB, YpeB, SubB, PltB, ArtB, *Ec*PltB, CtxB, and EcxB). A diverse sialoglycan binding pattern emerged when the phylogenetically-related AB_5 toxin B subunits were tested on the microarray and the dataset was organized using the coding system. The comparative binding affinity of the phylogenetically-related six B subunits is presented in a standard heatmap format (Fig. 9). A common observation made for all the B subunits (except for EcxB) under consideration was that they did not show any propensity towards the non-sialosides present in our library (SI Fig. *S2A*). Additionally, the B subunits failed to show any affinity towards glycans with a terminal Kdn (a rare mammalian Sia) unless there was an internal Neu5Ac/Gc present in the glycan structure (SI Fig. *S2B*).

A broad range of Sia-binding by YenB protein has already been demonstrated (Fig. 6). Interestingly, YpeB, a closely related B-subunit protein showed versatile Sia recognition albeit lacking the ability to recognize 4-*O*-acetylated Sia (Fig. 9). The inability of YpeB to bind Neu4,5Ac_2 and Neu4Ac5Gc may hold the key to answering why some mammals (like horses) are thought to be less susceptible to the plague. Neither of the B subunits could recognize the Kdn-glycans, except when the terminal Kdn is linked to an internal Neu5Ac or Neu5Gc in α2-8 linkage.

A previous report of glycan microarray analysis of *E. coli* SubB showed a stronger binding preference towards the glycans terminating with α2-3-linked Neu5Gc compared to its Neu5Ac congener (41). Testing of SubB in the expanded sialoglycan microarray revealed a similar binding pattern with a preferential α2-3-linked Neu5Gc binding, irrespective of the underlying glycan structures. A weaker albeit significant binding was observed towards α2-6 and α2-8-linked Neu5Gc. The modification on Neu5Gc showed an immense effect in SubB binding. A moderately strong binding was observed towards the 9-OAc Neu5Gc-glycans, however the 5-OMe and 4-OAc modified Neu5Gc completely eliminated the affinity towards SubB. The loss of ligandability of Neu5Gc upon 4-O-acetylation reaffirms the structural similarity of SubB to YpeB, which shares high sequence homology. Notably, Kdn associated glycans did not exhibit any binding to SubB.

Human specific typhoid toxin B subunit (PltB) binds exclusively with Neu5Ac-glycans (51). A detailed analysis of the binding preferences showed that the binding intensity decreases moderately when the terminal Neu5,9Ac_2 is linked to the underlying glycan in α2-6 linkage. Interestingly, the binding was completely diminished when the terminal Neu5Ac was protected with an 8-OMe group.

An interesting observation was made when ArtB was tested on the microarray. Relative to PltB, ArtB showed a broader Sia preference with strong propensity towards α2-3 and α2-8-linked Neu5Ac- and acetylated Neu5Ac-glycans. However, terminal Neu5Ac with α2-6-linkage exhibited 3 to 5-fold less binding intensity towards ArtB. Sialosides with terminal Neu5Gc showed a mild detrimental effect in binding with ArtB, which is unlike PltB where the Neu5Gc-glycan completely blocks Sia recognition.

*Ec*PltB has been reported to preferably bind with Neu5Ac-glycans when tested at a higher concentration (100μg/ml) but partially loses its selectivity at low concentration (1 μg/ml) (42). We tested the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein at 30 µg/ml concentration and a mild preference towards Neu5Ac over Neu5Gc-glycans was observed. However, the Sia linkage plays an important role in binding affinity. *Ec*PltB showed a clear preference towards α2-3 and α2-8-linked Sia and failed to bind with α2-6-linked Sia, a distinct trait that matches that seen in ArtB. Additionally, Sia recognition by *Ec*PltB was completely blocked by 4-OAc modified Neu5Ac/Gc; a similar observation was made for YpeB and SubB.

The glycan recognition by CtxB and EcxB was very different when compared to other toxin B subunits under consideration. It is known that CtxB preferentially binds to naturally occurring GM1, where the Sia presented is Neu5Ac (44). When tested on the sialoglycan microarray we found that CtxB distinctly recognized both Neu5Ac- and Neu5Gc-associated GM1, as well as fucosylated-GM1. As expected, the binding intensity greatly diminished with GD1 as a ligand. No other sialosides and non-sialosides showed any affinity towards CtxB.

Like CtxB, EcxB also displayed a strong binding preference towards Neu5Ac- and Neu5Gc-GM1. It also recognized the fucosylated-GM1 glycan, albeit with relatively weaker affinity. Like in CtxB, GD1 was also found to be a moderately weak receptor for EcxB. The most striking differences in glycan binding between CtxB and EcxB was the recognition of non-sialosides. Notably, EcxB but not CtxB recognized the non-sialosides present in our library, which represented the terminal fragments of glycosphingolipids of either lacto-, neo-lacto-, or ganglio-series (Fig. S2). Fig. 9 heatmap shows a comparison of Sia-binding preferences.

Application of the coding system to estimate the diversity of sialoglycans in nature: Calculation of possible sialyltrisaccharide structures. The systematic coding system also helped us calculate the number of possible sialyltrisaccharide structures that can exists in nature. In estimating the space of all, plausible and likely trisaccharides some fundamental concepts were applied. The set of *all* trisaccharides contains every trisaccharide code that can be written down. The set of *plausible* trisaccharides was extracted from that set by reasonably asserting that all sugars must be connected, and no modification or linkage can share a carbon; principles well established in glycobiology. Finally, the *likely* trisaccharides were extracted using a probabilistic approach to generalize glycan motifs found in a set of known trisaccharides. Many studies are now shifting attention to glycan motif discovery to bring generalizable meaning to lectin array analysis (14, 25, 52), glycan array analysis (24), glycan database analysis (53–55) and glycoprofile analysis (56). Functional meaning or biosynthetic feasibility discovered in the existence or enrichment of one motif can be used to create generalizable expectations regarding the functional importance or biosynthetic feasibility of a novel glycan (57). Here, we used co-occurrences observed between each adjacent and non-adjacent glycan element (sugar, linkage and modification) to determine which two-element motifs were common, observed or never observed. Similar to the studies using motifs to generalize structural feasibility, the joint probabilities were aggregated over each novel trisaccharide to calculate a motif-based likelihood for each plausible trisaccharide.

Enumeration of all plausible and likely trisaccharides.To enumerate all trisaccharide combinations and establish a total number for reference, we used a simple cartesian product; every selection of one element from each of several sets. Here, we enumerated every selection of one of the sugars, linkages or modifications permissible at each of the 9 positions. The cartesian product of the trisaccharide code gives 212,889,600 total trisaccharide combinations.

To calculate the number of plausible trisaccharides, we selected from the set of all trisaccharides by asserting that trisaccharides must be connected and non-conflicting. Intuitively, no nonreducing sugar may follow a missing linkage (i.e. “None”). Similarly, no linkage or modification may connect to a missing

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reducing monosaccharide (i.e. "None"). Yet, a missing modification has no impact on the plausibility of the trisaccharide. We also applied constraints specifying the specific carbons each sugar provides for the formation of a glycosidic bond and the carbons required by a linkage or modification. Each monosaccharide provides a discrete number of hydroxyl groups at specific carbons, while each modification and linkage requires a free hydroxyl group at specific carbons. In a plausible trisaccharide, each linkage and modification are associated with a sugar providing at least every necessary carbon to support those linkages and modifications. Additionally, the reducing linkage, modification, and nonreducing linkage must not require the same carbons. After removing all unconnected trisaccharides (those with an interceding "None"), all conflicting trisaccharides (those with linkages and modifications requiring the same or unavailable carbons), there are 2,799,962, only 1.32% of the total, plausible trisaccharides remaining.

Finally, we ranked the likelihood of trisaccharides by comparing them to a list of 140 known human trisaccharides. Because permissible reaction information is limited for trisaccharides, we used a statistical approach to infer feasible trisaccharide structures from the list of known human structures. We looked at the probability that any two elements in the trisaccharide code would co-occur in the set of 140 known sugars (9 choose 2 = 36 pairs) and used that to generate a log-likelihood scoring. Due to the limited coverage of the combination space, we added a pseudo-count (0.1) to minimize the penalty of pairs unobserved in the known set. Over every plausible trisaccharide, the median log-likelihood was -21.7 and the median number of unobserved pairs per plausible trisaccharide was 6. The least likely trisaccharides have low log-likelihoods (min=-70.9) and high numbers of unobserved pairs (max=34). There were 113,781 trisaccharides with no unobserved pairs and these least unusual trisaccharides had a median log-likelihood of -8.07. There are 471,292 trisaccharides with at most 2 unobserved pairs and 746,861 trisaccharides with at most 3 unobserved pairs. With more unobserved pairs, the likelihood of that trisaccharide naturally occurring decreases.

In counting the total, plausible and likely trisaccharides, we found that these numbers differed by several orders of magnitude. The number of total trisaccharides is useful to orient us to the size of combinatorial possibility but has limited biological meaning as it contains many impossible structures with, for example, 1st and 3rd sugars with no 2nd sugar. The possible trisaccharide structures impose the minimal constraint, the sugar must be connected (no missing links or sugars) and all carbons requested by linkages or modifications must be requested only once and supported by the corresponding sugar. These constraints reduced the total sugar set from over 210 million to just over 2 million. The plausible sugars provide the most forgiving possible constraints and therefore describe an upper-bound suggesting that there are unlikely to be more than 2.8 million trisaccharides. Finally, we imposed a harsh constraint, compliant with a set of known trisaccharides. 113,781 of the plausible trisaccharides were completely compliant with the set of known trisaccharides, containing no code pairs unobserved in the known set, suggesting that there are at least 113,781 sialyltrisaccharides.

Though an unobserved pair makes it less likely that a trisaccharide is sterically feasible to assemble, the set of known trisaccharides was not comprehensive. Therefore, a trisaccharide with an unobserved pair may simply be a new type of trisaccharide. The more uncharacterized pairs a trisaccharide has, the less likely it is to contain no impossible code pairs. Therefore, a more forgiving lower-bound could be established as fewer than 2 or 3 unobserved pairs, 471,000 and 747,000 trisaccharides respectively. Within these large steps, trisaccharides can be ranked by log-likelihood to find the most likely sugars within each set.

While we developed the system for the >113,000 plausible linear sialyltrisaccharides, we note that a biantennary *N*-glycan with two terminal sialoglycan trisaccharides could have $>10^{10}$ potential combinations and a triantennary *N*-glycan with three terminal sequences, $>10^{15}$ potential combinations at a single *N*-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

glycosylation site. With this calculation that includes selected sialic acid modifications and linkages, we extended the possibility of the huge number of N-glycans reported previously (58).

Conclusions and Perspectives

Sialoglycan microarray analysis presented an immense advantage in identifying novel interactions between sialoglycans and other biomolecules. We have established a chemoenzymatically synthesized defined sialyltrisaccharide library for the microarray technology platform. The challenges in sorting and finding motifs from vast microarray data were addressed by developing a novel 10-digit alpha-numerical coding system. The assignment of the code to the individual monosaccharide, their modification, and linkage made the analysis and visualization of data much easier and streamlined. This machine-readable code can be used to find binding target motifs. Using this coding system, we attempted to calculate the number of plausible sialyltrisaccharide structures in nature. Our calculation revealed a remarkable number of 10^5 possible linear structures of sialyltrisaccharides and hence gave an idea of the huge structural diversity in whole sialomes where the structures of post-translational glycan modifications are not always linear.

We evaluated the binding preferences of phylogenetically-related bacterial AB_5 toxin B subunits towards the sialoglycans using the sialoglycan microarray platform. We identified that the B subunit from *Y. enterocolitica* could recognize all Neu5Ac and Neu5Gc glycans present in our library irrespective of their modifications and linkages. This gives us an opportunity to use this toxin B subunit as a universal sialoglycan probe. We also reevaluated the sialoside binding affinity of other phylogenetically-related B subunits. They presented a diverse array (SI Table *S2*) of Sia-binding preferences, which can be exploited to develop a set of sialoglycan binding probes. We realize that it is impossible to have a microarray library that includes every possible naturally occuring sialoside. Moreover, the complexity of cell surface sialomes is beyond the scope of the defined sialoglycan library. To probe the immense diversity of sialosides in biological models, we propose the development of Sialoglycan Recognition Probes (SGRP).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Experimental Procedures

Expression and purification of *Yersenia enterocolitica* B-subunit (YenB)

The method described in accompanying paper (accompanying manuscript Khan *et al.*) was followed to express and purify YenB protein with a C-terminal His_6 tag. The codon-optimised open reading frame encoding YenB was chemically synthesized, flanked by an *Eco*RI site and a ribosome binding site 5' to the initiating ATG codon, and with a sequence encoding a His_6 tag fused to the C-terminus of the ORF and a 3' *Hin*dIII site, cloned into the respective sites of pUC57 (GenScript; Piscataway, NJ). The fragments were then excised and cloned into the *Eco*RI and *Hin*dIII sites of the expression vector pBAD18 (59) such that the ORF is under the control of the vector *ara* promoter. This construct was then transformed into *E. coli* BL21.

For protein purification, the recombinant bacterium was grown in 500 ml LB Broth supplemented with 50 µg/ml ampicillin at 37°C to late logarithmic phase, diluted in half with fresh medium supplemented with 0.2% arabinose to induce *yenB* expression, and then incubated overnight at 26°C. Cells were harvested by centrifugation, resuspended in 20 ml loading buffer (50 mM sodium phosphate, 300 mM NaCl, 20 mM imidazole, pH 8.0), and lysed in a French pressure cell. Cell debris was removed by centrifugation at 20,000 × *g* for 30 min at 4°C. The supernatant was then loaded onto a 2 ml column of Ni-nitrilotriacetic acid (NTA) resin, which had been preequilibrated with 20 ml loading buffer. The column was then washed with 40 ml loading buffer, and bound proteins were eluted with a 30-ml gradient of 0 to 500 mM imidazole in loading buffer; 3 ml fractions were collected and analyzed by SDS-PAGE, followed by staining with Coomassie blue or Western blotting with monoclonal anti-His_6. Fractions containing >95% pure YenB were pooled, concentrated and stored in PBS/50% glycerol at -15°C.

Modeling study of YenB with sialosides

TheYenB and other amino acid sequences were aligned using the program EXPRESSO (60), and homology modeling was performed employing MODELLER software (61), using the crystal structure of SubB (3DWP) as a template.

Sialoglycan microarray

The sialoglycan microarray method was adapted and modified from the literature reported earlier (29, 62). Defined sialosides with amine linker were chemoenzymatically synthesized and then quantitated utilizing an improved DMB-HPLC method (33). 100 µM of sialoglycan solution (in 300mM Na-phosphate buffer, pH 8.4) was printed in quadruplets on NHS-functionalized glass slides (PolyAn 3D-NHS; catalog# PO-10400401) using an ArrayIt SpotBot® Extreme instrument. The slides were blocked (0.05M ethanolamine solution in 0.1M Tris-HCl, pH 9.0), washed with warm Milli-Q water and dried.

The slides were then fitted in a multi-well microarray hybridization cassette (ArrayIt, CA) to divide into 8 wells and rehydrated with 400 µl of Ovalbumin (1% w/v, PBS) for one hour in a humid chamber with gentle shaking. After that the blocking solution was discarded and a 400 µl of solution of the toxin B-subunit (30 µg/ml) in the same blocking buffer was added to the individual well. The slides were incubated for 2h at room temperature with gentle shaking and the slides were washed with PBS-Tween (0.1% v/v). The wells were then treated with Cy3-conjugated anti-His (Rockland Antibodies & Assays; Cat# 200-304-382) secondary antibody at 1:500 dilution in PBS. Followed by gentle shaking for 1 hour in the dark and humid chamber. The slides were then washed, dried, and scanned with a Genepix 4000B scanner (Molecular Devices Corp., Union City, CA) at wavelength 532 nm. Data analysis was performed using the Genepix Pro 7.3 software (Molecular

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Devices Corp., Union City, CA). The raw data analysis and sorting using the numerical codes were performed on Microsoft Excel.

Enumeration and estimation of trisaccharides

In estimating the space of all plausible and likely trisaccharides some fundamental concepts were applied. The set of *all* trisaccharides contains every trisaccharide code that can be written down. The set of *plausible* trisaccharides was extracted from that set by reasonably asserting that all sugars must be connected, and no modification or linkage can share a carbon; principles well established in glycobiology. Finally, the *likely* trisaccharides were extracted using a probabilistic approach to generalize glycan motifs found in a set of known trisaccharides.

In a plausible trisaccharide, each linkage and modification are associated with a sugar providing at least every necessary carbon to support those linkages and modifications. Additionally, the reducing linkage, modification, and nonreducing linkage must not require the same carbons. Intuitively, no nonreducing sugar may follow a missing linkage (i.e. "None"). Similarly, no linkage or modification may connect to a missing reducing monosaccharide (i.e. "None"). Yet, a missing modification has no impact on the plausibility of the trisaccharide. We also applied constraints specifying the specific carbons each sugar provides for the formation of a glycosidic bond and the carbons required by a linkage or modification. Each monosaccharide provides a discrete number of hydroxyl groups at specific carbons, while each modification and linkage requires a free hydroxyl group at specific carbons.

To calculate likely trisaccharides, We looked at the probability that any two elements in the trisaccharide code would co-occur in the set of 140 known sugars (9 choose 2 = 36 pairs) and used that to generate a log-likelihood scoring. Due to the limited coverage of the combination space, we added a pseudo-count (0.1) to minimize the penalty of pairs unobserved in the known set.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

This work was supported by NIH Grant R01GM32373 (to A.V.), the Novo Nordisk Foundation (NNF10CC1016517, NNF20SA0066621), and NIGMS R35GM119850 (to N.E.L).

Conflicts of Interest

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1: Coding strategy utilized to encode sialyltrisaccharide. Codes associated with the individual glycan identity, modification, and linkage are listed starting from the non-reducing end. Example of 10-digit numerical code representation of sialyltrisaccharides. New glycan id (extreme right column) is generated based on the encoding rules described in Table 1. The left most column describes the semi random old glycan id which is based on the order of which the glycans were synthesized and stored. For complete list see SI Table *S1*.

Figure 2: Application of coding system to identify binding pattern. Chicken polyclonal anti-Neu5Gc IgY antibody was tested on sialoglycan microarray platform and average fluorescence unit (Ave RFU) was plotted against a) semi-random glycan id and b) new glycan code arranged according to the terminal Sia identity. The lone strong signal at the left of the figure (B) corresponds to a Neu5Ac-glycan with a branched Neu5Gc.

Figure 3: Application of coding system for logical sorting of microarray data. The data obtained from the microarray experiment with biotinylated SNA lectin was organized based on the linkage between the terminal Sia and the underlying glycan structure. The plot conveniently and clearly shows the preference of SNA lectin towards the α2,6-linked Sia.

Figure 4: Motif finding utilizing the new sialoglycan code. Top 20 strongest signal obtained from sialoglycan microarray experiment of Biotinylated Maackia Amurensis Lectin II (MAL II) was listed along with the 10-digit new glycan id. The digits labelled in red showing a clear motif in glycan structure that serves as the ligand epitope of MAL II. The most prominent motif determined here is the Siaα3Galβ3GalNAcα i.e, O-glycan type motif.

Figure 5: Phylogenetic relationship between species based on 16SrRNA vs pentavalent B subunit of exotoxins. The comparison reveals that the evolution of bacterial AB_5 toxin B subunit is independent of the species phylogeny. Corresponding accession numbers are indicated below.

Figure 6: Sialoglycan microarray experiment with B subunit from *Y. enterocolitica*. The chart showed broad range sialic acid recognition by the His-tagged YenB protein (30µg/ml). The non-sialoside and Kdn-glycan epitopes were not recognized by the YenB. Ganglioside type glycan epitopes are designated as asterix symbol (*) in the chart.

Figure 7: YenB structure modelling and comparison with SubB and YpeB. A) Superposition of the SubB X-ray structure (pale pink) (PDB code:3DWP) with model of YenB (cyan). Overall, very similar to SubB, YenB has longer helix at a distance from the binding site (showed by arrows). B) Superposition of the YpeB (pale green) and YenB (cyan) 3-dimensional model. The structure models are very similar to each other except a few longer helices at a distance from the binding site (showed by arrows).

Figure 8: Interaction of sialic acid (Neu5Gc) at the binding pocket of YenB and others. A) Superposition of the Neu5Gc binding site of SubB (pale pink) with the model of YenB (cyan) using the C-alpha coordinates of all residues. Neu5Gc is highlighted in light purple. B) Superposition of the Neu5Gc (light purple) binding

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

site of SubB (pink) with the model of YpeB (pale green) and model of YenB (cyan) using the C-alpha coordinates of all residues. Neu5Gc is highlighted in light purple.

Figure 9: Differential sialoglycan recognition by phylogenetically-related toxin B subunits. Data obtained from sialoglycan microarray experiment showed a diverse Sia binding affinity. The glycan structures were grouped according to the terminal Sia, modifications, and linkages to identify the binding pattern of various phylogenetically-related B subunits. For a qualitative binding comparison see supp. Table *S2*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Numerical code for designating naturally occurring trisaccharide sequences carrying terminal sialic acids. In first digit place alphabetical codes are used for less common or bacterial terminal glycan sequences carrying nonulosonic acid units.

Code	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
	Terminal Sia/NulO	**Mod.**	**Link**	**Second Sugar**	**Mod/ Branch**	**Link**	**Third sugar**	**Mod/ Branch**	**Link**	**Additional**
1	Neu5Ac	1,7 Lactone		Neu5Ac	Fucα1-2	α1-3	Neu5Ac	(Neu5Acα2-6)	α1-3	Gal
2	Neu4,5Ac$_2$	(8OS)		Gal	Galα1-3	β1-3	Gal	(GalNAcβ1-4)	β1-3	Glcβ1
3	Neu5,9Ac$_2$	9-O-Lactyl	α2-3		(3OS)	α2-3		(GlcNAcβ1-6)	α2-3	GlcNAcβ1
4	Neu5Gc	4,9-di-OAc	α2-4	Neu5Gc	(GalNAcβ1-4)	β1-4	Neu5Gc	(Neu5Gcα2-6)	β1-4	Man
5	Neu4Ac5Gc	5-OMe	α2-5				Glc	(Fucα1-4)	β1-2	
6	Neu5Gc9Ac	7,8,9-triOAc	α2-6	GalNAc	(6OS)	α2-6	GalNAc	(6OS)	α2-6	Galβ1-4Glcβ1
7	Kdn	7-OAc	α2-7	Kdn			Kdn	(Fucα1-3)	β1-6	Galβ1-4GlcNAc
8	Kdn4Ac	8-OMe	α2-8	GlcNAc		α2-8	GlcNAc	(6OS, Fucα1-3)	α2-8	α1-1' (with linker)
9	Kdn9Ac	(7/8)9-diOAc	α2-9			α2-9		(6OS, Fucα1-4)		β1-1' (with linker)
0	None	None	None	None	None	None	None	None	None	
a	Neu	1,5 Lactam		Neu	(Neu5Acα2-3-)		Neu	(Galβ1-4GlcNAcβ1-6)		
b	Leg				(Neu5Gcα2-3-)			(Neu5Acα2-3Galβ1-4GlcNAcβ1-6)		
c	Leg5,7Ac$_2$				(Kdnα2-3-)			(Neu5Acα2-6Galβ1-4GlcNAcβ1-6)		
d	Pse									
e	Aci				(Galβ1-3GalNAcβ1-4)			(Galβ1-3GalNAcβ1-4)		
f	4eLeg5,7Ac$_2$							(Neu5Acα2-3Galβ1-3GalNAcβ1-4)		
g	8eLeg5,7Ac$_2$							(Neu5Acα2-6Galβ1-3GalNAcβ1-4)		
h					(Fucα1-2Galβ1-3GalNAcβ1-4)					

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Varki, A., Schnaar, R. L. and Schauer, R. (2017) in *Essentials of Glycobiology* (Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Darvill, A. G., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L. and Seeberger, P. H., eds.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY)
2. Varki, A. and Angata, T. (2006) Siglecs--the major subfamily of I-type lectins. *Glycobiology* **16**, 1R-27R
3. Pérez, S. and Mulloy, B. (2005) Prospects for glycoinformatics. *Curr Opin Struct Biol* **15**, 517-524
4. Von der Lieth, C. W., Bohne-Lang, A., Lohmann, K. K. and Frank, M. (2004) Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform* **5**, 164-178
5. Laine, R. A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4**, 759-767
6. von der Lieth, C. (2004) An endorsement to create open access databases for analytical data of complex carbohydrates. *Journal of carbohydrate chemistry* **23**, 277-297
7. McNaught, A. D. (1997) International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology. Joint Commission on Biochemical Nomenclature. Nomenclature of carbohydrates. *Carbohydr Res* **297**, 1-92
8. Bohne-Lang, A., Lang, E., Förster, T. and von der Lieth, C. W. (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* **336**, 1-11
9. Banin, E., Neuberger, Y., Altshuler, Y., Halevi, A., Inbar, O., Nir, D. and Dukler, A. (2002) A novel linear code® nomenclature for complex carbohydrates. *Trends in Glycoscience and Glycotechnology* **14**, 127-137
10. Shinmachi, D., Yamada, I., Aoki, N. P., Matsubara, M., Aoki-Kinoshita, K. F. and Narimatsu, H. (2017) in *A Practical Guide to Using Glycomics Databases* eds.) pp. 41-73, Springer,
11. Sahoo, S. S., Thomas, C., Sheth, A., Henson, C. and York, W. S. (2005) GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr Res* **340**, 2802-2807
12. Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y. and Narimatsu, H. (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics* **21**, 1717-1718
13. Kellman, B. P., Zhang, Y., Logomasini, E., Meinhardt, E., Godinez-Macias, K. P., Chiang, A. W. T., Sorrentino, J. T., Liang, C., Bao, B., Zhou, Y., Akase, S., Sogabe, I., Kouka, T., Winzeler, E. A., Wilson, I. B. H., Campbell, M. P., Neelamegham, S., Krambeck, F. J., Aoki-Kinoshita, K. F. and Lewis, N. E. (2020) A consensus-based and readable extension of Linear Code for Reaction Rules (LiCoRR). *Beilstein J Org Chem* **16**, 2645-2662
14. Klamer, Z., Staal, B., Prudden, A. R., Liu, L., Smith, D. F., Boons, G. J. and Haab, B. (2017) Mining High-Complexity Motifs in Glycans: A New Language To Uncover the Fine Specificities of Lectins and Glycosidases. *Anal Chem* **89**, 12342-12350
15. Doubet, S., Bock, K., Smith, D., Darvill, A. and Albersheim, P. (1989) The Complex Carbohydrate Structure Database. *Trends Biochem Sci* **14**, 475-477
16. Doubet, S. and Albersheim, P. (1992) CarbBank. *Glycobiology* **2**, 505
17. Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., Ueda, N., Hamajima, M., Kawasaki, T. and Kanehisa, M. (2005) KEGG as a glycome informatics resource. *Glycobiology*
18. Raman, R., Venkataraman, M., Ramakrishnan, S., Lang, W., Raguram, S. and Sasisekharan, R. (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* **16**, 82R-90R
19. Toukach, P., Joshi, H. J., Ranzinger, R., Knirel, Y. and von der Lieth, C. W. (2007) Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res* **35**, D280-6

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R. D., York, W. S., Karlsson, N. G., Lisacek, F., Packer, N. H., Campbell, M. P., Aoki, N. P., Fujita, A., Matsubara, M., Shinmachi, D., Tsuchiya, S., Yamada, I., Pierce, M., Ranzinger, R., Narimatsu, H. and Aoki-Kinoshita, K. F. (2017) GlyTouCan: an accessible glycan structure repository. *Glycobiology* **27**, 915-919
21. Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N. and Kornfeld, S. (2015) Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **25**, 1323-1324
22. Cao, Y., Park, S. J., Mehta, A. Y., Cummings, R. D. and Im, W. (2020) GlyMDB: Glycan Microarray Database and analysis toolset. *Bioinformatics* **36**, 2438-2442
23. Padler-Karavani, V., Song, X., Yu, H., Hurtado-Ziola, N., Huang, S., Muthana, S., Chokhawala, H. A., Cheng, J., Verhagen, A., Langereis, M. A., Kleene, R., Schachner, M., de Groot, R. J., Lasanajak, Y., Matsuda, H., Schwab, R., Chen, X., Smith, D. F., Cummings, R. D. and Varki, A. (2012) Cross-comparison of protein recognition of sialic acid diversity on two novel sialoglycan microarrays. *J Biol Chem* **287**, 22593-22608
24. Cholleti, S. R., Agravat, S., Morris, T., Saltz, J. H., Song, X., Cummings, R. D. and Smith, D. F. (2012) Automated motif discovery from glycan array data. *OMICS* **16**, 497-512
25. Tang, H., Hsueh, P., Kletter, D., Bern, M. and Haab, B. (2015) The detection and discovery of glycan motifs in biological samples using lectins and antibodies: new methods and opportunities. *Adv Cancer Res* **126**, 167-202
26. Coff, L., Chan, J., Ramsland, P. A. and Guy, A. J. (2020) Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics* **21**, 42
27. Deng, L., Chen, X. and Varki, A. (2013) Exploration of sialic acid diversity and biology using sialoglycan microarrays. *Biopolymers* **99**, 650-665
28. Bensing, B. A., Khedri, Z., Deng, L., Yu, H., Prakobphol, A., Fisher, S. J., Chen, X., Iverson, T. M., Varki, A. and Sullam, P. M. (2016) Novel aspects of sialoglycan recognition by the Siglec-like domains of streptococcal SRR glycoproteins. *Glycobiology* **26**, 1222-1234
29. Lu, N., Ye, J., Cheng, J., Sasmal, A., Liu, C. C., Yao, W., Yan, J., Khan, N., Yi, W., Varki, A. and Cao, H. (2019) Redox-Controlled Site-Specific α2-6-Sialylation. *J Am Chem Soc* **141**, 4547-4552
30. Yusufi, F. N., Park, W., Lee, M. M. and Lee, D. Y. (2009) An alpha-numeric code for representing N-linked glycan structures in secreted glycoproteins. *Bioprocess Biosyst Eng* **32**, 97-107
31. Chen, X. and Varki, A. (2010) Advances in the biology and chemistry of sialic acids. *ACS Chem Biol* **5**, 163-176
32. Ye, J., Liu, X.-w., Peng, P., Yi, W., Chen, X., Wang, F. and Cao, H. (2016) Diversity-Oriented Enzymatic Modular Assembly of ABO Histo-blood Group Antigens. *ACS Catal.* **6**, 8140-8144
33. Ji, Y., Sasmal, A., Li, W., Oh, L., Srivastava, S., Hargett, A. A., Wasik, B. R., Yu, H., Diaz, S., Choudhury, B., Parrish, C. R., Freedberg, D. I., Wang, L. P., Varki, A. and Chen, X. (2021) Reversible O-Acetyl Migration within the Sialic Acid Side Chain and Its Influence on Protein Recognition. *ACS Chem Biol*
34. O'Loughlin, E. V., Gall, D. G. and Pai, C. H. (1990) Yersinia enterocolitica: mechanisms of microbial pathogenesis and pathophysiology of diarrhoea. *J Gastroenterol Hepatol* **5**, 173-179
35. Toma, S. and Lafleur, L. (1974) Survey on the incidence of Yersinia enterocolitica infection in Canada. *Appl Microbiol* **28**, 469-473
36. Black, R. E., Jackson, R. J., Tsai, T., Medvesky, M., Shayegani, M., Feeley, J. C., MacLeod, K. I. and Wakelee, A. M. (1978) Epidemic Yersinia enterocolitica infection due to contaminated chocolate milk. *N Engl J Med* **298**, 76-79
37. Kay, B. A., Wachsmuth, K., Gemski, P., Feeley, J. C., Quan, T. J. and Brenner, D. J. (1983) Virulence and phenotypic characterization of Yersinia enterocolitica isolated from humans in the United States. *J Clin Microbiol* **17**, 128-138

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Wang, X., Cui, Z., Wang, H., Tang, L., Yang, J., Gu, L., Jin, D., Luo, L., Qiu, H., Xiao, Y., Xiong, H., Kan, B., Xu, J. and Jing, H. (2010) Pathogenic strains of Yersinia enterocolitica isolated from domestic dogs (Canis familiaris) belonging to farmers are of the same subtype as pathogenic Y. enterocolitica strains isolated from humans and may be a source of human infection in Jiangsu Province, China. *J Clin Microbiol* **48**, 1604-1610
39. Mantle, M. and Husar, S. D. (1993) Adhesion of Yersinia enterocolitica to purified rabbit and human intestinal mucin. *Infect Immun* **61**, 2340-2346
40. Mantle, M. and Husar, S. D. (1994) Binding of Yersinia enterocolitica to purified, native small intestinal mucins from rabbits and humans involves interactions with the mucin carbohydrate moiety. *Infect Immun* **62**, 1219-1227
41. Byres, E., Paton, A. W., Paton, J. C., Löfling, J. C., Smith, D. F., Wilce, M. C., Talbot, U. M., Chong, D. C., Yu, H., Huang, S., Chen, X., Varki, N. M., Varki, A., Rossjohn, J. and Beddoe, T. (2008) Incorporation of a non-human glycan mediates human susceptibility to a bacterial toxin. *Nature* **456**, 648-652
42. Littler, D. R., Ang, S. Y., Moriel, D. G., Kocan, M., Kleifeld, O., Johnson, M. D., Tran, M. T., Paton, A. W., Paton, J. C., Summers, R. J., Schembri, M. A., Rossjohn, J. and Beddoe, T. (2017) Structure-function analyses of a pertussis-like toxin from pathogenic Escherichia coli reveal a distinct mechanism of inhibition of trimeric G-proteins. *J Biol Chem* **292**, 15143-15158
43. Kaper, J. B., Morris, J. G. and Levine, M. M. (1995) Cholera. *Clin Microbiol Rev* **8**, 48-86
44. De Haan, L. and Hirst, T. R. (2004) Cholera toxin: a paradigm for multi-functional engagement of cellular mechanisms (Review). *Mol Membr Biol* **21**, 77-92
45. Alisson-Silva, F., Liu, J. Z., Diaz, S. L., Deng, L., Gareau, M. G., Marchelletta, R., Chen, X., Nizet, V., Varki, N., Barrett, K. E. and Varki, A. (2018) Human evolutionary loss of epithelial Neu5Gc expression and species-specific susceptibility to cholera. *PLoS Pathog* **14**, e1007133
46. Cervin, J., Wands, A. M., Casselbrant, A., Wu, H., Krishnamurthy, S., Cvjetkovic, A., Estelius, J., Dedic, B., Sethi, A., Wallom, K. L., Riise, R., Bäckström, M., Wallenius, V., Platt, F. M., Lebens, M., Teneberg, S., Fändriks, L., Kohler, J. J. and Yrlid, U. (2018) GM1 ganglioside-independent intoxication by Cholera toxin. *PLoS Pathog* **14**, e1006862
47. Wands, A. M., Fujita, A., McCombs, J. E., Cervin, J., Dedic, B., Rodriguez, A. C., Nischan, N., Bond, M. R., Mettlen, M., Trudgian, D. C., Lemoff, A., Quiding-Järbrink, M., Gustavsson, B., Steentoft, C., Clausen, H., Mirzaei, H., Teneberg, S., Yrlid, U. and Kohler, J. J. (2015) Fucosylation and protein glycosylation create functional receptors for cholera toxin. *Elife* **4**, e09545
48. Karasawa, T., Ito, H., Tsukamoto, T., Yamasaki, S., Kurazono, H., Faruque, S. M., Nair, G. B., Nishibuchi, M. and Takeda, Y. (2002) Cloning and characterization of genes encoding homologues of the B subunit of cholera toxin and the Escherichia coli heat-labile enterotoxin from clinical isolates of Citrobacter freundii and E. coli. *Infect Immun* **70**, 7153-7155
49. Ng, N. M., Littler, D. R., Paton, A. W., Le Nours, J., Rossjohn, J., Paton, J. C. and Beddoe, T. (2013) EcxAB is a founding member of a new family of metalloprotease AB5 toxins with a hybrid cholera-like B subunit. *Structure* **21**, 2003-2013
50. Dhar, C., Sasmal, A. and Varki, A. (2019) From “Serum Sickness” to “Xenosialitis”: Past, Present, and Future Significance of the Non-human Sialic Acid Neu5Gc. *Front Immunol* **10**, 807
51. Deng, L., Song, J., Gao, X., Wang, J., Yu, H., Chen, X., Varki, N., Naito-Matsui, Y., Galán, J. E. and Varki, A. (2014) Host adaptation of a bacterial toxin from the human pathogen Salmonella Typhi. *Cell* **159**, 1290-1299
52. Klamer, Z., Hsueh, P., Ayala-Talavera, D. and Haab, B. (2019) Deciphering Protein Glycosylation by Computational Integration of On-chip Profiling, Glycan-array Data, and Mass Spectrometry. *Mol Cell Proteomics* **18**, 28-40
53. Kellman, B. P. and Lewis, N. E. (2021) Big-Data Glycomics: Tools to Connect Glycan Biosynthesis to Extracellular Communication. *Trends Biochem Sci* **46**, 284-300

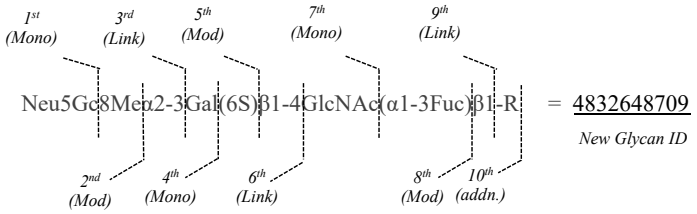
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

54. Hosoda, M., Takahashi, Y., Shiota, M., Shinmachi, D., Inomoto, R., Higashimoto, S. and Aoki-Kinoshita, K. F. (2018) MCAW-DB: A glycan profile database capturing the ambiguity of glycan recognition patterns. *Carbohydr Res* **464**, 44-56
55. Klein, J. and Zaia, J. (2019) glypy: An Open Source Glycoinformatics Library. *J Proteome Res* **18**, 3532-3537
56. Benedetti, E., Pučić-Baković, M., Keser, T., Wahl, A., Hassinen, A., Yang, J. Y., Liu, L., Trbojević-Akmačić, I., Razdorov, G., Štambuk, J., Klarić, L., Ugrina, I., Selman, M. H. J., Wuhrer, M., Rudan, I., Polasek, O., Hayward, C., Grallert, H., Strauch, K., Peters, A., Meitinger, T., Gieger, C., Vilaj, M., Boons, G. J., Moremen, K. W., Ovchinnikova, T., Bovin, N., Kellokumpu, S., Theis, F. J., Lauc, G. and Krumsiek, J. (2017) Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway. *Nat Commun* **8**, 1483
57. Bojar, D., Powers, R. K., Camacho, D. M. and Collins, J. J. (2021) Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions. *Cell Host Microbe* **29**, 132-144.e3
58. Cummings, R. D. (2009) The repertoire of glycan determinants in the human glycome. *Mol Biosyst* **5**, 1087-1104
59. Guzman, L. M., Belin, D., Carson, M. J. and Beckwith, J. (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* **177**, 4121-4130
60. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. and Notredame, C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* **340**, 385-395
61. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. and Sali, A. (2006) Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **Chapter 5**, Unit-5.6
62. Meng, C., Sasmal, A., Zhang, Y., Gao, T., Liu, C. C., Khan, N., Varki, A., Wang, F. and Cao, H. (2018) Chemoenzymatic Assembly of Mammalian O-Mannose Glycans. *Angew Chem Int Ed Engl* **57**, 9003-9007

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

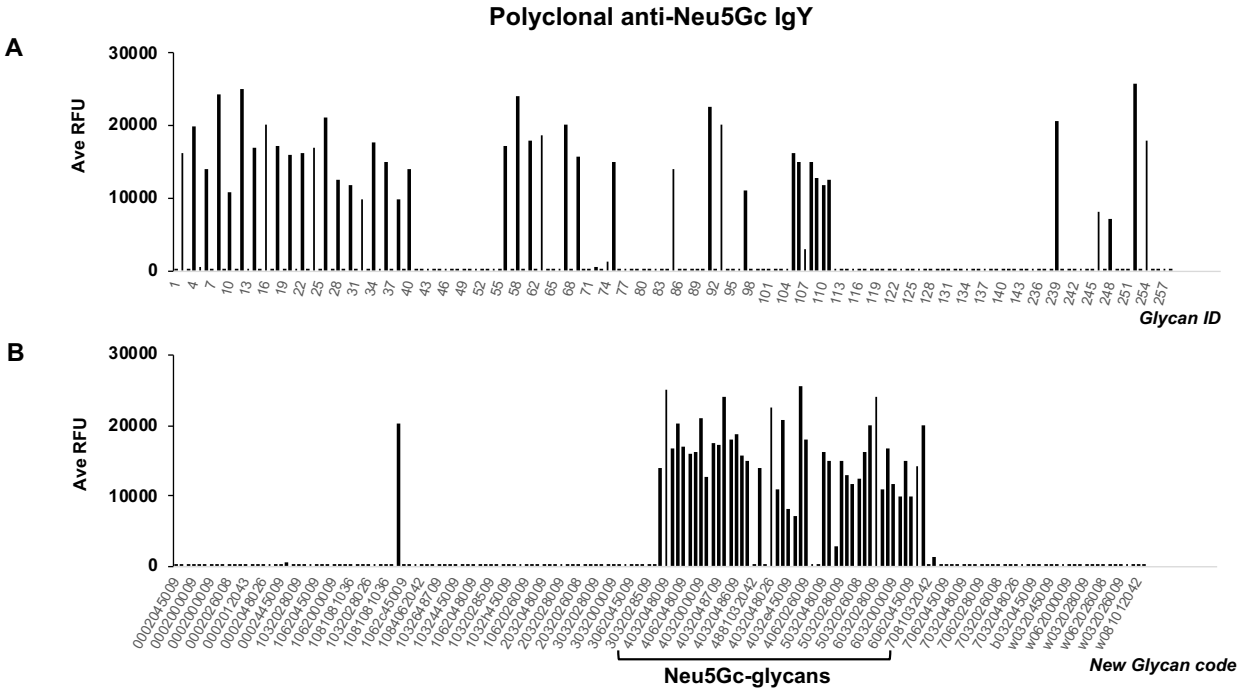
Fig. 1 *Sasmal et al.*



Glycan ID	Glycan Structure	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	New Glycan ID
1	Neu5,9Ac2α3Galβ4GlcNAcβR1	3	0	3	2	0	4	8	0	0	9	3032048009
2	Neu5Gc9Acα3Galβ4GlcNAcβR1	6	0	3	2	0	4	8	0	0	9	6032048009
3	Neu5,9Ac2α6Galβ4GlcNAcβR1	3	0	6	2	0	4	8	0	0	9	3062048009
4	Neu5Gc9Acα6Galβ4GlcNAcβR1	6	0	6	2	0	4	8	0	0	9	6062048009
5	Neu5Acα6GalNAcαR1	1	0	6	6	0	0	0	0	0	8	1066000008
6	Neu5Gcα6GalNAcαR1	4	0	6	6	0	0	0	0	0	8	4066000008
7	Neu5,9Ac2α3Galβ3GlcNAcβR1	3	0	3	2	0	2	8	0	0	9	3032028009
8	Neu5Gc9Acα3Galβ3GlcNAcβR1	6	0	3	2	0	2	8	0	0	9	6032028009
9	Neu5,9Ac2α3Galβ3GalNAcαR1	3	0	3	2	0	2	6	0	0	8	3032026008
10	Neu5Gc9Acα3Galβ3GalNAcαR1	6	0	3	2	0	2	6	0	0	8	6032026008

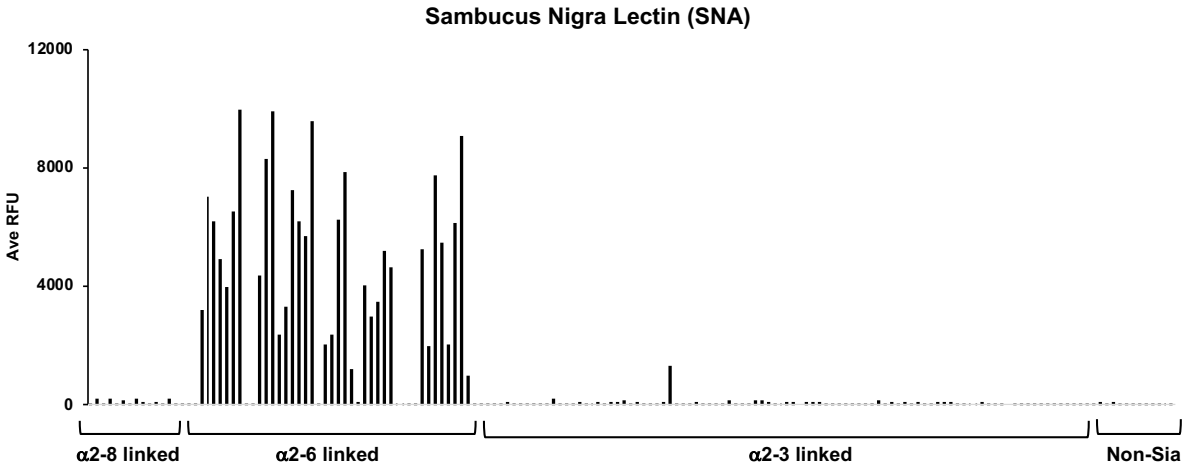
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 2 *Sasmal et al.*



bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3 *Sasmal et al.*



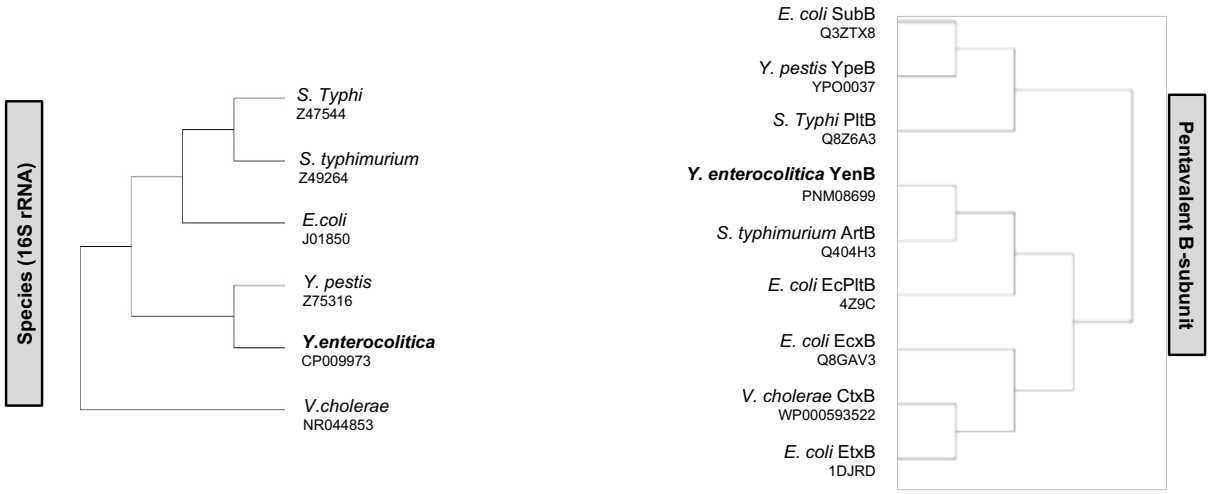
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 4 *Sasmal et al.*

Glycan Structure	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	New Glycan ID	Ave RFU
Neu5Gcα3Galβ3GalNAcαR1	4	0	3	2	0	2	6	0	0	8	40 32026 008	
Neu5,9Ac2α3Galβ3GalNAcαR1	3	0	3	2	0	2	6	0	0	8	30 32026 008	
Neu5Acα3Galβ3GalNAcαR1	1	0	3	2	0	2	6	0	0	8	10 32026 008	
Neu4Ac5Gcα3Galβ3GalNAcαR1	5	0	3	2	0	2	6	0	0	8	50 32026 008	
Neu5Acα3Galβ3GalNAcβR1	1	0	3	2	0	2	6	0	0	9	10 32026 009	
Neu4Ac5Gcα3Galβ3GlcNAcαR1	5	0	3	2	0	2	8	0	0	8	50 32028 008	
Neu5Gcα3Galβ3GalNAcβR1	4	0	3	2	0	2	6	0	0	9	40 32026 009	
Neu4,5Ac2α3Galβ3GalNAcαR1	2	0	3	2	0	2	6	0	0	8	20 32026 008	
Neu4,5Ac2α3Galβ3GalNAcβR1	2	0	3	2	0	2	6	0	0	9	20 32026 009	
Neu4Ac5Gcα3Galβ3GalNAcβR1	5	0	3	2	0	2	6	0	0	9	50 32026 009	
Neu4,5Ac2α3Galβ3GlcNAcβR1	2	0	3	2	0	2	8	0	0	9	20 32028 009	
Neu4,5Ac2α-3Galβ3GlcNAcαR1	2	0	3	2	0	2	8	0	0	8	20 32028 008	
Kdnα3Galβ3GalNAcαR1	7	0	3	2	0	2	6	0	0	8	70 32026 008	
Neu5Gc9Acα3Galβ3GalNAcαR1	6	0	3	2	0	2	6	0	0	8	60 32026 008	
Neu5Acα3Galβ3GlcNAcαR1	1	0	3	2	0	2	8	0	0	8	10 32028 008	
Neu4Ac5Gcα3Galβ3GlcNAcβR1	5	0	3	2	0	2	8	0	0	9	50 32028 009	
Neu5,9Ac2α3Galβ3GalNAcβR1	3	0	3	2	0	2	6	0	0	9	30 32026 009	
Neu5Gc9Acα3Galβ3GalNAcβR1	6	0	3	2	0	2	6	0	0	9	60 32026 009	
Neu5Gcα3Galβ3GlcNAcβR1	4	0	3	2	0	2	8	0	0	9	40 32028 009	

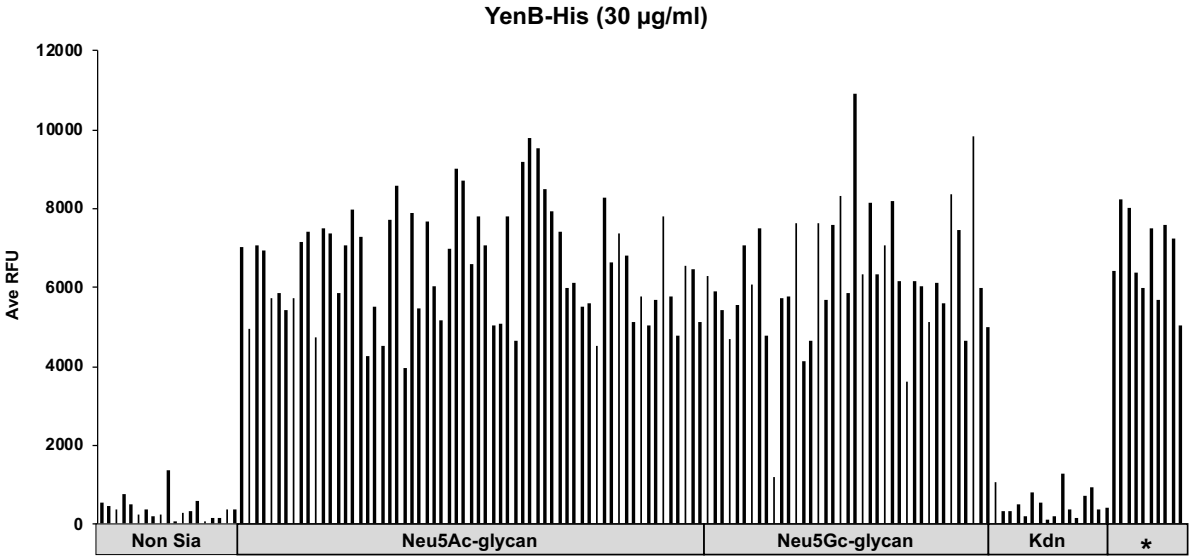
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 5 *Sasmal et al.*



bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 6 *Sasmal et al.*



bioRxiv preprint doi: https://doi.org/10.1101/2021.05.28.446191; this version posted May 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 7 *Sasmal et al.*

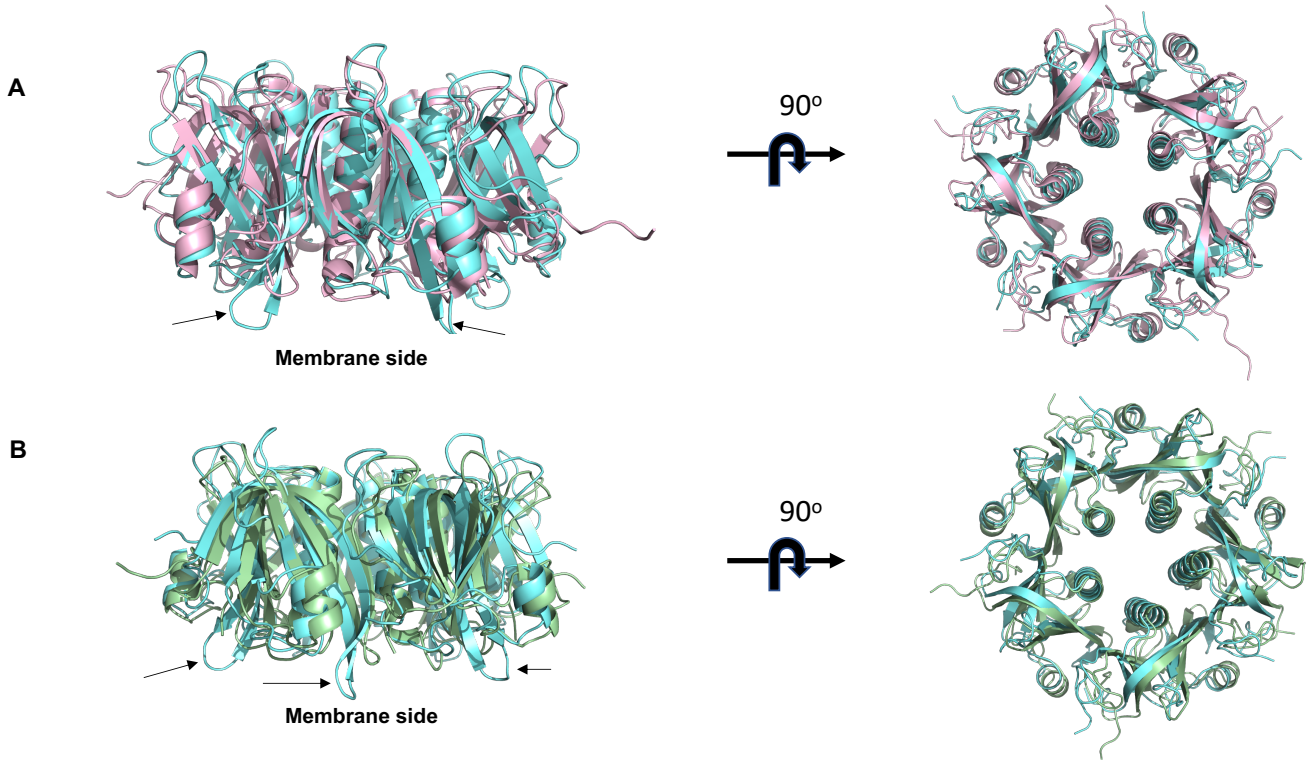


Fig. 8 *Sasmal et al*.

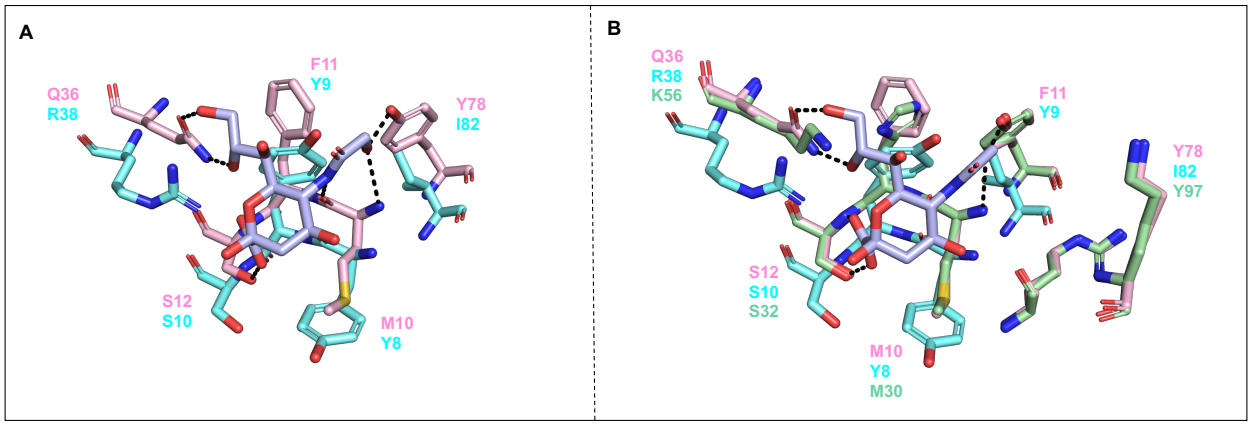
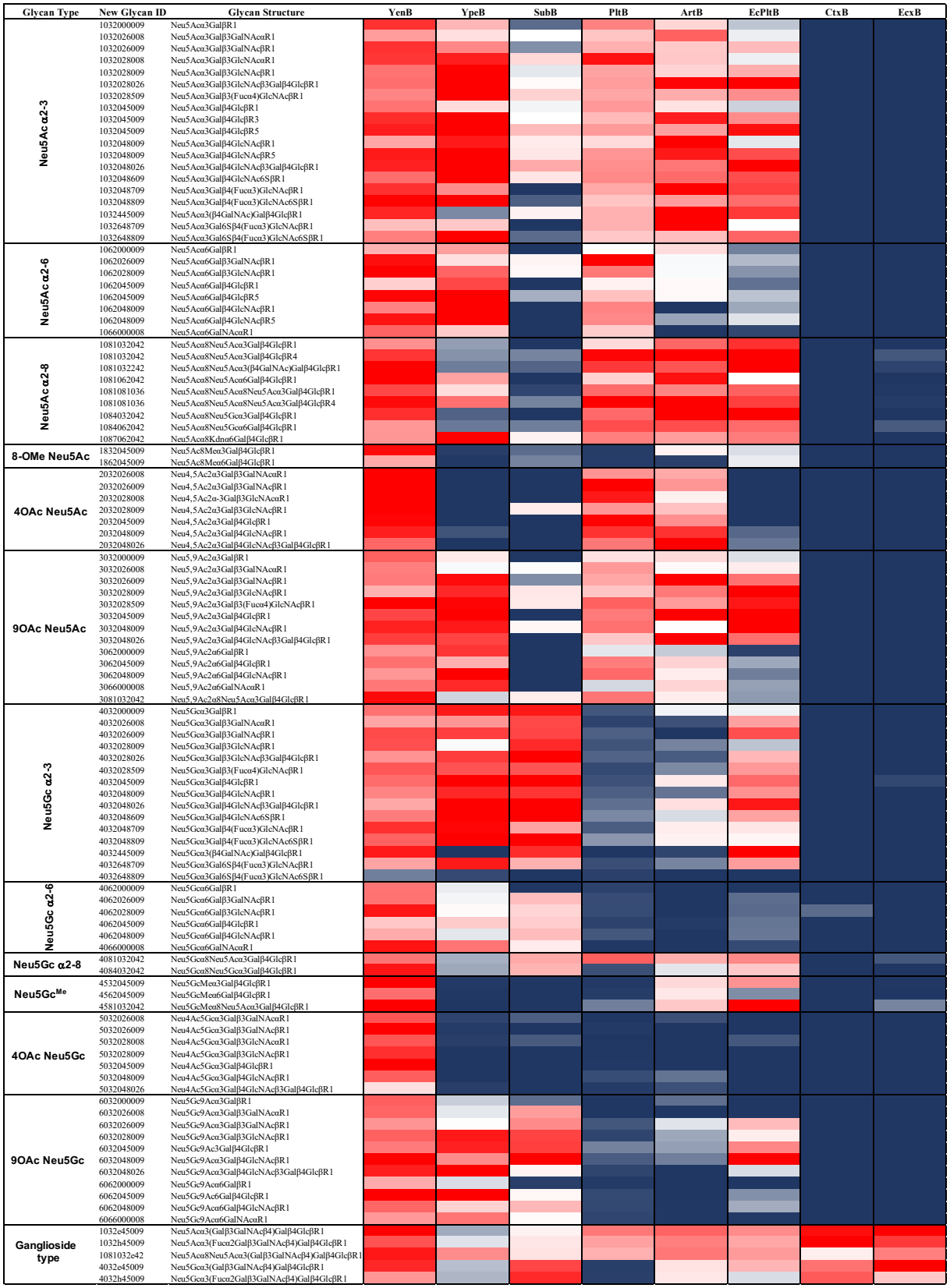



Fig. 9 *Sasmal et al.*