Title

CarboGrove: a resource of glycan-binding specificities through analyzed glycan-array datasets from all platforms

Authors

Zachary L. Klamer¹, Chelsea M. Harris², Jonathan M. Beirne², Jessica E. Kelly², Jian Zhang², and Brian B. Haab^{1*}

Affiliations

¹Van Andel Institute, Grand Rapids, MI

²Z Biotech, Aurora, CO

*brian.haab@vai.org

Abstract

The volume and value of glycan-array data are increasing, but no common method and resource exists to analyze, integrate, and use the available data. To meet this need, we developed a resource of analyzed glycan-array data called CarboGrove. Building on the ability to process and interpret data from any type of glycan array, we populated the database with the results from 35 types of glycan arrays, 13 glycan families, 5 experimental methods, and 19 laboratories or companies. In meta-analyses of glycan-binding proteins, we observed glycan-binding specificities that were not uncovered from single sources. In addition, we confirmed the ability to efficiently optimize selections of glycan-binding proteins to be used in experiments for discriminating between closely related motifs. CarboGrove yields unprecedented access to the wealth of glycan-array data being produced and powerful capabilities for both experimentalists and bioinformaticians.

Teaser

We introduce a resource that allows researchers to find, compare, study, and integrate analyses from all types of glycan-array data.

37 Introduction

38 Glycan arrays are being produced and used by more labs than ever before. After the first
39 reports of glycan arrays in 2004 (1, 2), just a handful of laboratories worked on the
40 technology for about the next decade. The main provider of the technology to the research
41 community was the Consortium for Functional Glycomics (CFG). The CFG array, which
42 used the planar-array method that had been established for DNA and protein arrays,
43 contained glycans that represented a broad survey of the known, important motifs in
44 mammalian biology. The significance of the CFG resource was that it provided access to
45 researchers who were not able to produce arrays themselves, primarily by reason of the
46 cost and difficulty of synthesizing glycans. The array was heavily used, and it was the
47 large number of experiments performed on this platform that established the value of the
48 technology and stimulated further developments in the field, including in experimental
49 methods, bioinformatics tools, and methods of glycan synthesis. The CFG data were the
50 sole source used for multiple bioinformatics efforts in the analysis of glycan array data (3–
51 6).

52 But no single array could include all the glycans necessary for every study. The diversity
53 and number of structures present among various classes of glycans and organism types is
54 too great even for the largest array. This situation drove researchers to develop arrays with
55 content organized around specific fields of research. For example, plant biologists and
56 microbial biologists each developed glycan arrays relevant to their fields (7, 8), and
57 researchers studying sialic acids developed arrays with a wide range of variants of that
58 feature (9). Other features of specialized content include glycosaminoglycans (10, 11),
59 glycopeptides (12–14), and human milk oligosaccharides (15). A major hurdle in each
60 setting was the production of glycans, as each glycan requires unique synthesis and
61 purification, but this limitation could be addressed through improved synthetic strategies
62 and automated synthesizers (16, 17), in combination with focused production around
63 specific types of glycans. Furthermore, researchers have developed experimental
64 alternatives to the planar array. The novel technologies—methods involving mass
65 spectrometry (18) or bacteriophage display (19), for example—provide complementary
66 information and capabilities to the planar array and allow dispersion of the methods to a
67 greater number of laboratories.

68 All of these developments have resulted in an expanding variety of glycan-array data
69 available for study. Bioinformatics methods that could capture and use all available
70 glycan-array data, regardless of source and content, could serve many purposes, from
71 learning more about the specificity of a particular protein, to finding lectins with a pre-
72 defined specificity, to larger-scale, integrative studies in glycobiology. Such analyses
73 could not be done manually, given the complexity of glycans and protein-glycan
74 interactions, as well as the complexity of integrating information over many data points
75 from the array. The data from the various sources must be analyzed and interpreted with a
76 common system.

77 Several resources currently provide glycan-array data in either raw or analyzed form:
78 CFG, LFDB (20), MCAW-DB (21), and GlyMDB (6). These resources represent valuable
79 advances in the field, but they have limited value as a general resource for non-
80 bioinformaticians. One limitation is that each provides data for only a single type of array,
81 either CFG (CFG, MCAW-DB, and GlyMDB) or frontal-affinity chromatography
82 (LFDB). Further, the resources provide powerful visualization tools, such as the GLAD
83 (GLycan Array Dashboard) employed by GlyMDB, but limited or incomplete
84 interpretation of the data. For example, MCAW-DB gives an alignment of top-binding
85 glycans, which gives clues about features associated with binding, but it does not provide

the context of their binding strength relative to others, which is necessary to achieve a complete picture. In general, the discernment of the specificity of a glycan-binding protein requires significant, additional analysis on the part of the researcher. The need for algorithms to discern the complex, fine specificities of glycan-binding proteins is clear, given the sensitivity of binding to minor differences in glycans such as the position of the epitope on N-glycan branches (22, 23).

Given software to reliably interpret data from any platform, a resource could be built that provides common access to glycan-array information across the many sources that are now available. We recently introduced the MotifFinder software (24, 25) to meet the analysis need. We demonstrated earlier that the algorithm delivers a detailed and accurate analysis of the specificity of a glycan-binding protein, and that it can perform the analyses through the integration of data from distinct platforms. These previous developments suggested an approach to unify the analysis and usage of glycan array data. In the present work, we asked whether a database system driven by the MotifFinder engine could meet the need for a unified glycan-array resource.

Results

Achieving common data processing across platforms and array types

The available glycan-array data cover several approaches to detection and quantification (Fig. 1A). This diversity represents a challenge when collating data into a common platform, but it also provides complementary information from the strengths and limitations of each platform. The planar array uses robust methods that were established for DNA arrays and thus has been a workhorse for many labs, but it has the limitation that non-specific or reduced binding that can occur from the linker (26), the surface (27), or the tagging of the glycan-binding protein (28). In addition, most embodiments do not account for glycan density or kinetics. The newer technologies require specialized methods or equipment, but they provide solution-phase kinetics (18, 19), incorporation of density as a controllable parameter (14), or display on a cell-surface context (29).

The available data also contain a great diversity of glycans (Fig. 1B). The CFG array was heavily weighted toward mammalian glycans, but technology developers have branched out into microbial, plant cell-wall, and other non-mammalian glycans. In addition, improved synthesis technologies have resulted in glycans with increased complexity and a broader variety of monosaccharides, as well as a variety of glycosaminoglycans. These developments have resulted in an increased frequency in reports on new glycan arrays (Fig. 1C and Supplemental Table 1).

We sought to develop a system that provides a common mode of analysis for all available data. To account for the diversity in glycans, we developed a parser that translates text representations of all types of monosaccharide names and their connections into a common format used by the program. To enable processing of data from any type of array or platform, we developed the analysis algorithm to be independent of scale or range but require only a quantitative value corresponding to each glycan in the array positively associated with binding.

The various glycan-array data could then be processed in our MotifFinder algorithm for identifying the motifs—patterns within glycans—that best describe the specificity of the glycan-binding protein applied to the array. The algorithm uses data from multiple concentrations of the protein, if available, to give more accurate results than possible from one concentration (24) (Fig. 2A). The family of motifs that defines the specificity (referred to as the model) is arranged into two types: the primary motifs, which represent distinct

structural categories, and the fine-specificity motifs, which represent gradations in binding within the primary motifs (Fig. 2B). The model is visualized in various ways to assist user interpretation (Fig. 2C). The consistent output across all datasets, regardless of platform or type of glycans, is a critical component of enabling cross-dataset comparisons and searches.

We tested the ability of the algorithm to process and organize glycan-array data from 35 different types of arrays, 13 different glycan families, 5 experimental methods, and 19 laboratories or companies (Fig. 3A and Supplemental Table 2). These included publicly available data as well as unpublished data (Supplemental Tables 2 and 3). The number of contributions from each provider ranged from 1 to 541 datasets, for a total of 1125 datasets (Supplemental Table 4). MotifFinder was able to produce a model for each of the glycan-binding proteins with only minor adjustments in formatting required for some datasets.

We then assembled the models into a relational database called CarboGrove (Fig. 3A). This collection promises to cover a much broader range of glycans and the glycan-binding proteins than any single resource. To evaluate the scope of the database, we defined 118 different motifs from 11 families based on a set obtained from the GlyGen resource (30) with the addition of motifs covering the major core types (Supplemental Tables 5 and 6). This list is not exhaustive but provides an initial, unbiased survey of the breadth of the database. An analysis of the glycans on the arrays showed that all motifs were represented on at least one array, and that some were on nearly every array (Fig. 3B). The arrays had a broader range of inclusion of motifs, ranging from only 1 to 112, reflecting the variation in purposes of the arrays.

To determine whether the glycan-binding proteins in the database cover a broad range of specificities, we used the model for each protein to predict binding to each of 1803 glycans that spanned all arrays. From the resulting values, we assessed binding to each of the 118 motifs using a motif score (24). All but 19 of the 118 motifs are bound (motif score > 2) by 4 or more glycan-binding proteins (Fig. 3C and Supplemental Table 5). The 19 not bound by any proteins represent less-common features such as type 3 A antigen. Very common motifs such as N-Glycan and Biantennary N-Glycan are bound by many glycan-binding proteins: 258 and 208, respectively. These differences reflect both the prevalence of the motif in biology as well as the amount of research centered on the motif.

Accessing and analyzing glycan arrays across platforms

The collection of analyzed data potentially offers access to detailed, accurate information about the specificity of any given glycan-binding protein. We sought to enable such searches through a system of matching user-specified terms with all relevant datasets. This task involved accounting for variability in the conventions in common names, abbreviations, and the use of the terms lectin and agglutinin. To address this difficulty, we included multiple aliases for each protein and allowed relevant results to be returned even when a search does not match the primary name in the database (Supplemental Methods).

We tested the search and analysis capabilities using the lectin SNA (*Sambucus nigra* agglutinin). The primary specificity of SNA, α 2,6-linked sialic acids, is well known, but the fine specificities are not well understood owing to limited variety in glycans containing α 2,6-linked sialic acid on the arrays and the complexity in the analysis. A search for SNA returned 36 individual datasets from 16 sources (Fig. 4). Nearly every source confirmed the canonical specificity of SNA. The top motifs for each array, however, revealed complementary information. The ASN-Linked and AsymmetricN

arrays, which focus on N-linked glycans, identified a preference for the tri/tetra-antennary presentation (motifs A1 and A2, ASN-Linked array) over the biantennary presentation (motif A0, ASN-Linked array), as well as a preference for the 3' mannose branch (motif A0, AsymmetricN array) over the 6' mannose branch or unbranched presentations (motif B0, AsymmetricN array). The CFG array, which has the greatest diversity in the α 2,6-sialyl-LacNAc motif, identified preferential binding on extended N-linked glycans over O-linked glycans (motifs A0 and A4). Some arrays had limited variation in the α 2,6-sialyl epitope and consequently produced ambiguous, incomplete motifs, such as motifs B0 and D0 on the Gildersleeve array and A0 and the NGGM-TM, NGGM, and CEMA O-GalNAc arrays. Arrays with a low reliability score (as assessed by the amount of nonbinding-glycan noise), such as the LiGA and GlycanBead arrays, generated various weak motifs that are inconsistent with results from other arrays.

We also tested this functionality on the lectin wheat germ agglutinin (WGA), which is widely used but has a specificity that is poorly understood. Part of the challenge is its breadth in specificity, binding nearly half of the glycans on the CFG array (273/609 glycans) when applied at high concentrations. The search gave results from 12 different arrays (Supplemental Figure 1). A comparative analysis identified both known and novel features, such as highest binding to 6' linked terminal GlcNAc and 3' linked Neu5Ac; binding to both GlcNAc and GalNAc in other terminal linkages, provided the 3' carbon is unsubstituted; and potential binding to the heparan sulfate motif GlcNAc α 1-4GlcA. The novel observations would require experimental confirmation, but they are structurally plausible and demonstrate findings that are made possible through broad analyses of glycan-array data.

Selecting glycan-binding proteins for experimental design

A companion capability is to specify a motif and to search for glycan-binding proteins that bind the motif. We selected three motifs for a test of this function: N-glycan core fucose, Lewis X, and type-2 blood group B. These motifs have the common feature of fucose, but they differ in the fucose linkage: either to the 6', 3', or 2' carbon of the adjoining monosaccharide. A hierarchical cluster of all the models in the database and the set of motifs defined above indicated that the search motifs are bound by separate groups of proteins (Fig. 5A). The top 10 glycan-binding proteins for each motif confirmed that each motif returned a unique set of proteins known to bind the motif (Table 1). An assessment of the top motifs bound by the glycan-binding proteins showed that each protein is a specific binder of the search motif (Fig. 5B).

A closer analysis showed the value of an unbiased search. For example, the top hits for motif "N-Glycan α 6 Core Fucose" did not include the lectin commonly used for this motif, Lens Culinaris Agglutinin (LCA), because LCA bound many N-glycans without α -6 fucose (as shown in the CarboGrove model, not shown). The Lewis X and blood group B searches likewise returned results that were unexpected but potentially useful, such as the strong binding to Lewis X of both the anti-Lewis X antibody and the lectins LTA, CCL2 and AAL.

This functionality suggested an additional opportunity in experimental design. Searches such as demonstrated above could be modified to optimally select glycan-binding proteins for an experiment, such as to distinguish between motifs in a biological sample that are difficult to distinguish by mass spectrometry. We tested this concept for terminal N-acetyl-galactosamine (GalNAc), in either the α and β orientation, and terminal N-acetyl-glucosamine (GlcNAc), in either the bisecting or outer-arm position. These features are isomers but have important differences in biological function.

We sought to identify a minimal set of lectins (limited to 3-4) that would give optimal distinction between the comparison motifs. First, for each of the four terminal features (alpha-GalNAc, beta-GalNAc, outer-arm GlcNAc, and bisecting GlcNAc), we defined glycans containing the feature (Fig. 6A). We also defined negative-control glycans that have the core structures but not the terminal features. Next, we searched the database to identify lectins that bind any of the motifs. We predicted the binding of each lectin to each glycan and assembled the values (Fig. 6A), from which we could search for combinations of lectins that give unique patterns of binding across each of the comparison motifs. Multiple algorithms are available for maximizing distances between subsets. For demonstration, we used manually guided optimization to arrive at the minimal set of GSL-II, HAA, VVL, and PHA-E (Fig. 6B). The average binding of the lectins to the glycans in the comparison groups showed distinct patterns, corresponding to the differences in the top motifs (Fig. 6B): GSL-II binds non-bisecting terminal GlcNAc; HAA binds terminal alpha-GalNAc; VVL binds LacDiNAc and some lipid-linked glycans; and PHA-E binds bisecting GlcNAc. Thus, starting from the full collection of >700 models, we efficiently reduced to just 4 that provide clear distinctions among the 4 isomeric motifs.

Discussion

The proliferation of glycan-array platforms and data has precipitated a need for a common mode of analyzing, interpreting, and accessing the data. Here we provide a solution via the MotifFinder analysis program and the CarboGrove database. We populated the database with analyzed data from 35 types of arrays and from multiple suppliers and platforms. Thus, for the first time, researchers can access and use an expansive collection of glycan-array data in an analyzed form. The ability to bring together data from multiple sources is especially important in the case of glycan arrays, where each type of array provides information or experiments that are complementary to the others. In particular, these arrays are complementary in their glycans, the glycan-binding proteins analyzed, and the strengths and weaknesses of their experimental systems. The enabling component of this project was a software tool for analyzing all types of glycan-array data. Without a common model of analyzing and reporting the data, the assembly and integrated analysis of such diverse data would be prohibitively time-consuming and inaccurate.

The ability to compare and integrate results between separate datasets has advantages for many applications. In the evaluation of SNA, for which the search returned 36 datasets, we observed fine specificities that are distinct from the canonical specificity and that would be missed in the evaluation of only a single platform. The findings were consistent with the previous studies of SNA. For example, the MotifFinder identification of preferential binding to the primary epitope on extended N-linked glycans from the 3' mannose branch agrees with the manual analysis (22). But the ability to conveniently supplement the study with comparable findings from other arrays, such as the preference for tri/tetra-antennary N-glycans over linear presentations (Fig. 4) is unique.

These analyses also expand on previous methods to compare between platforms. Previous studies to compare between conditions or platforms generally used manual analyses, such as comparisons between arrays that focused on sialic acids (31) or between experimental conditions (22, 32). An algorithm-based approach was introduced in a well-designed study to compare results between separate array platforms (33). The authors used a universal thresholding method to identify differences between platforms that correlated with experimental differences. The method did not, however, capture both weak and strong binding for all arrays and did not account for fact that no single concentration was relevant in comparing results between arrays. In contrast, the algorithm used in the present work is

sensitive to weakly bound glycans and allows comparisons among arrays even where concentrations are not optimally matched. Furthermore, the software has the unique support of combined analyses of multiple datasets, previously demonstrated for integrating data from multiple concentrations of a lectin or from multiple platforms (24).

Besides studying the specificities of glycan-binding proteins, a major function of CarboGrove is to identify glycan-binding proteins that have user-specified binding traits. This function will be important for both non-experts and experts in glycobiology. Searches across lesser-known platforms and motifs can be impractical, and the specificities of some glycan-binding proteins are not generally known or are misunderstood. But even more valuable could be the ability to select optimal sets of lectins for experiments. The use of lectins to detect or quantify glycan structures is very common, for example in methods such immunofluorescence, Western blot, cell-staining, in-vivo imaging, and others, but in many cases, the experiments do not employ the optimal lectins or are inaccurately interpreted. Using CarboGrove, a researcher could perform searches to identify a limited number of glycan-binding proteins that target the motifs that are relevant to the biological study. MotifFinder could predict binding to a set of relevant glycans and select the proteins most useful for analysis. Bioinformaticians could support this work by developing tools employing additional approaches to optimize experiments that use glycan-binding proteins.

The current study and resource have several limitations. The analysis algorithm does not account for certain features that could influence binding, such as the method of attaching the glycan, the density of the glycan, or the nature of a polypeptide backbone if present. A wide variety of experimental conditions also can influence apparent binding, including buffer, array substrate, and detection method (28). The resource currently does not house much of the metadata associated with the experiments, from which one could explore factors that influence binding. The metadata exists in a great variety of completeness and form, so the inclusion of all available information in the database is a technical hurdle. The database also does not house the raw data, which could be a drawback for researchers interested in deeper analyses. The database is not intended to be a resource for data deposition, but conceivably such a development could be useful. Finally, enabling support for user-defined motifs for database searches could expand the specificity-finding capabilities of the resource. This addition requires additional developments in the motif-building tools and would dramatically increase the computational overhead of the database and thereby database operating costs, one of the limiting factors in bioinformatics-resource lifecycles.

The pace of introduction of new arrays and glycans is clearly quickening. Versatile systems of attachment to surfaces using both covalent and non-covalent deposition (27) could make array production easier for non-specialist labs. Others have displayed glycans on the surface of bacteriophage (19), produced various glycans through sequential knockdowns of genes in glycan biosynthesis (29), and tuned the density of glycans spotted on chip through efficient methods of producing glycopeptide arrays (14). Thus, additional types of information from the new technologies could be accessible and analyzed. The MotifFinder platform supports updating to allow for additional factors to be explored. Our ongoing work involves support for investigating the influence of peptide backbone, glycan density, and linker type, as well as experimental factors that have been shown to introduce variability (32).

In addition to expansions of the current system, we foresee separate opportunities for informatics developers arising from CarboGrove. Developers could use the models in CarboGrove for downstream glycan-analysis applications, for example, as when a

researcher probes a biological sample with one or more glycan-binding proteins. The researcher could enter the quantified binding data into a program that uses the reports from CarboGrove to interpret the data. We previously demonstrated the feasibility and value of this approach (34). With further development, one could use the motif syntax in MotifFinder as the link to integrate lectin-binding data with data from mass spectrometry. Or in purely computational studies, bioinformaticians could use the resource to explore relationships among families of lectins and antibodies in association with genetic or organismal information. In support of such work, the motifs from MotifFinder could be used as the connection to a wide range of data on sequence, biosynthesis, and other information, such as are accessible through the GlyGen resource (30) and other databases.

Materials and Methods

Experimental Design

The objective of the study was to determine if a common method of analyzing glycan-array data could provide the contents for a comprehensive resource to retrieve, integrate, and use results from glycan arrays. To evaluate this question, we tested the ability of the system to analyze data from any available dataset; to provide broad coverage of motifs and glycan-binding proteins than available from single sources; to allow meta-analyses across datasets that give novel insights into fine specificities; and to efficiently select glycan-binding proteins that are optimal for a given experiment.

Data Collection

The collaborators at Z Biotech provided 151 glycan-array datasets generated for internal quality control studies, including data for 43 different glycan-binding proteins collected on 11 of the arrays offered by the company. Raw data are available as supplemental data (Supplemental Table 3). Details on the data collection are provided in the supplemental MIRAGE document (Supplemental Table 2).

Data for the CFG and NCFG arrays were retrieved from their respective websites and databases. Data from individual laboratories were retrieved from the original publications or provided by the authors upon request.

Statistical Analysis

The prediction of binding to glycans using the models was described previously (24) and are details in the user's manual provided with MotifFinder. The Reliability Score, which was used to rank models by the quality of their results (Fig. 4), measures the difference between the average binding of the top motif and the average binding of the non-binding motifs, normalized by the standard deviation in the non-binding motifs. This metric is similar to the signal-to-noise ratio. Given the average of the top-binding glycan values m , the average of the non-binding glycan values v , the standard deviation of the non-binding glycan values s , and the number of datasets in the model n , the reliability score is calculated as:

$$reliability\ score = \frac{\sum_1^n (\max(m_n) - v_n) / s_n}{n}$$

Data Analysis

The majority of datasets could be analyzed as they were provided, with minor corrections in text syntax to match the modified IUPAC syntax used by the CFG. In one case, the CUPRA array, the data needed to be inverted to enforce the requirement of a positive

association between binding and the quantification of the binding. All data were analyzed using MotifFinder release version V2.2.5 (35)

The Motif Score was calculated as described previously (25). Briefly, a t-test with unequal variances is performed comparing the predicted binding of glycans with the motif to those without the motif. The resulting p-value is log10 transformed and re-signed to match the sign of the t value from the t-test. Motif Score values are truncated to the range of -10 to 10. The Motif Score is used to rank associations with binding rather than for statistics. Additional metrics are used to break ties in the Motif Score for motif sorting, including the average binding of glycans with the motif and the precision of the motif (the number of glycans with the motif that have a positive predicted binding divided by the number of glycans with the motif).

Database Design

The database was developed using MariaDB (an open-source MySQL branch) and delivered using php (version 7.1.28) for server-side processing of searches and database interface. Javascript was used for the web interface and the compression of webserver file uploads. Bash scripts and the jq json parser tool were used to process, manage, and return webserver requests. The database and webserver are hosted using Amazon Web Services.

References

1. O. Blixt, S. Head, T. Mondala, C. Scanlan, M. E. Huflejt, R. Alvarez, M. C. Bryan, F. Fazio, D. Calarese, J. Stevens, N. Razi, D. J. Stevens, J. J. Skehel, I. van Die, D. R. Burton, I. A. Wilson, R. Cummings, N. Bovin, C.-H. Wong, J. C. Paulson, Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *P Natl Acad Sci Usa*. 101, 17033–17038 (2004).
2. M. C. Bryan, F. Fazio, H.-K. Lee, C.-Y. Huang, A. Chang, M. D. Best, D. A. Calarese, O. Blixt, J. C. Paulson, D. Burton, I. A. Wilson, C.-H. Wong, Covalent Display of Oligosaccharide Arrays in Microtiter Plates. *J Am Chem Soc*. 126, 8640–8641 (2004).
3. S. R. Cholleti, S. Agravat, T. Morris, J. H. Saltz, X. Song, R. D. Cummings, D. F. Smith, Automated Motif Discovery from Glycan Array Data. *Omics J Integr Biology*. 16, 497–512 (2012).
4. L. Coff, J. Chan, P. A. Ramsland, A. J. Guy, Identifying Glycan Motifs Using a Novel Subtree Mining Approach. *Bmc Bioinformatics*. 21, 42 (2020).
5. A. Porter, T. Yue, L. Heeringa, S. Day, E. Suh, B. B. Haab, A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins. *Glycobiology*. 20, 369–380 (2010).
6. Y. Cao, S.-J. Park, A. Y. Mehta, R. D. Cummings, W. Im, GlyMDB: Glycan Microarray Database and Analysis Toolset. *Bioinformatics* (2019), doi:10.1093/bioinformatics/btz934.
7. C. Ruprecht, M. P. Bartetzko, D. Senf, P. Dallabernadina, I. Boos, M. C. F. Andersen, T. Kotake, J. P. Knox, M. G. Hahn, M. H. Clausen, F. Pfrengle, A Synthetic Glycan Microarray Enables Epitope Mapping of Plant Cell Wall Glycan-Directed Antibodies. *Plant Physiol*. 175, 1094–1104 (2017).
8. A. Geissner, A. Reinhardt, C. Rademacher, T. Johannssen, J. Monteiro, B. Lepenies, M. Thépaut, F. Fieschi, J. Mrázková, M. Wimmerova, F. Schuhmacher, S. Götze, D. Grünstein, X. Guo, H. S. Hahm, J. Kandasamy, D. Leonori, C. E. Martin, S. G. Parameswarappa, S. Pasari, M. K. Schlegel, H. Tanaka, G. Xiao, Y. Yang, C. L. Pereira, C. Anish, P. H. Seeberger, Microbe-focused glycan array screening platform. *Proc National Acad Sci*. 116, 201800853 (2019).
9. X. Song, H. Yu, X. Chen, Y. Lasanajak, M. M. Tappert, G. M. Air, V. K. Tiwari, H. Cao, H. A. Chokhawala, H. Zheng, R. D. Cummings, D. F. Smith, A Sialylated Glycan Microarray Reveals Novel Interactions of Modified Sialic Acids with Proteins and Viruses. *J Biol Chem*. 286, 31610–31622 (2011).
10. M. Horton, G. Su, L. Yi, Z. Wang, Y. Xu, V. Pagadala, F. Zhang, D. A. Zaharoff, K. Pearce, R. J. Linhardt, J. Liu, Construction of heparan sulfate microarray for investigating the binding of specific saccharide sequences to proteins. *Glycobiology*. 31, 188–199 (2020).
11. P. Chopra, A. Joshi, J. Wu, W. Lu, T. Yadavalli, M. A. Wolfert, D. Shukla, J. Zaia, G.-J. Boons, The 3-O-sulfation of heparan sulfate modulates protein binding and lyase degradation. *Proc National Acad Sci*. 118, e2012935118 (2021).
12. H. Hinou, S. Kikuchi, R. Ochi, K. Igarashi, W. Takada, S.-I. Nishimura, Synthetic glycopeptides reveal specific binding pattern and conformational change at O-

mannosylated position of α -dystroglycan by POMGnT1 catalyzed GlcNAc modification. *Bioorgan Med Chem.* 27, 2822–2831 (2019).

- 13. R. Nason, C. Büll, A. Konstantinidi, L. Sun, Z. Ye, A. Halim, W. Du, D. M. Sørensen, F. Durbesson, S. Furukawa, U. Mandel, H. J. Joshi, L. A. Dworkin, L. Hansen, L. David, T. M. Iverson, B. A. Bensing, P. M. Sullam, A. Varki, E. de Vries, C. A. M. de Haan, R. Vincentelli, B. Henrissat, S. Y. Vakhrushev, H. Clausen, Y. Narimatsu, Display of the human mucinome with defined O-glycans by gene engineered cells. *Nat Commun.* 12, 4070 (2021).**
- 14. M. Mende, A. Tsouka, J. Heidepriem, G. Paris, D. S. Mattes, S. Eickelmann, V. Bordoni, R. Wawrzinek, F. F. Fuchsberger, P. H. Seeberger, C. Rademacher, M. Delbianco, A. Mallagaray, F. F. Loeffler, On-Chip Neo-Glycopeptide Synthesis for Multivalent Glycan Presentation. *Chem European J.* 26, 9954–9963 (2020).**
- 15. A. R. Prudden, L. Liu, C. J. Capicciotti, M. A. Wolfert, S. Wang, Z. Gao, L. Meng, K. W. Moremen, G.-J. Boons, Synthesis of asymmetrical multiantennary human milk oligosaccharides. *Proc National Acad Sci.* 114, 6954–6959 (2017).**
- 16. J. Zhang, C. Chen, M. R. Gadi, C. Gibbons, Y. Guo, X. Cao, G. Edmunds, S. Wang, D. Liu, J. Yu, L. Wen, P. G. Wang, Machine-Driven Enzymatic Oligosaccharide Synthesis by Using a Peptide Synthesizer. *Angewandte Chemie Int Ed.* 57, 16638–16642 (2018).**
- 17. T. Li, L. Liu, N. Wei, J.-Y. Yang, D. G. Chapla, K. W. Moremen, G.-J. Boons, An automated platform for the enzyme-mediated assembly of complex oligosaccharides. *Nat Chem.* 11, 229–236 (2019).**
- 18. P. I. Kitov, E. N. Kitova, L. Han, Z. Li, J. Jung, E. Rodrigues, C. D. Hunter, C. W. Cairo, M. S. Macauley, J. S. Klassen, A Quantitative, High-Throughput Method Identifies Protein–Glycan Interactions via Mass Spectrometry. *Commun Biology.* 2, 268 (2019).**
- 19. M. Sojitra, S. Sarkar, J. Maghera, E. Rodrigues, E. J. Carpenter, S. Seth, D. F. Vinals, N. J. Bennett, R. Reddy, A. Khalil, X. Xue, M. R. Bell, R. B. Zheng, P. Zhang, C. Nycholat, J. J. Bailey, C.-C. Ling, T. L. Lowary, J. C. Paulson, M. S. Macauley, R. Derda, Genetically encoded multivalent liquid glycan array displayed on M13 bacteriophage. *Nat Chem Biol.* 1–11 (2021).**
- 20. J. Hirabayashi, H. Tateno, T. Shikanai, K. F. Aoki-Kinoshita, H. Narimatsu, The Lectin Frontier Database (LfDB), and Data Generation Based on Frontal Affinity Chromatography. *Molecules.* 20, 951–973 (2015).**
- 21. M. Hosoda, Y. Takahashi, M. Shiota, D. Shinmachi, R. Inomoto, S. Higashimoto, K. F. Aoki-Kinoshita, MCAW-DB: A Glycan Profile Database Capturing the Ambiguity of Glycan Recognition Patterns. *Carbohydr Res.* 464, 44–56 (2018).**
- 22. L. Li, W. Guan, G. Zhang, Z. Wu, H. Yu, X. Chen, P. G. Wang, Microarray Analyses of Closely Related Glycoforms Reveal Different Accessibilities of Glycan Determinants on N-Glycan Branches. *Glycobiology.* 30, 334–345 (2019).**
- 23. S. Wang, C. Chen, M. Guan, D. Liu, X.-F. Wan, L. Li, Terminal Epitope-Dependent Branch Preference of Siglecs Toward N-Glycans. *Frontiers Mol Biosci.* 8, 645999 (2021).**
- 24. Z. Klammer, B. Haab, Combined Analysis of Multiple Glycan-Array Datasets: New Explorations of Protein–Glycan Interactions. *Anal Chem.* 93, 10925–10933 (2021).**

25. Z. Klamer, B. Staal, A. R. Prudden, L. Liu, D. F. Smith, G.-J. Boons, B. Haab, Mining High-Complexity Motifs in Glycans: A New Language To Uncover the Fine Specificities of Lectins and Glycosidases. *Anal Chem.* 89, 12342–12350 (2017).
26. O. C. Grant, H. M. Smith, D. Firsova, E. Fadda, R. J. Woods, Presentation, presentation, presentation! Molecular-level insight into linker effects on glycan array screening data. *Glycobiology.* 24, 17–25 (2014).
27. C. Li, A. S. Palma, P. Zhang, Y. Zhang, C. Gao, L. M. Silva, Z. Li, F. Trovão, M. Weishaupt, P. H. Seeberger, L. M. Likhoshesterov, V. Piskarev, J. Yu, U. Westerlind, W. Chai, Non-Covalent Microarrays from Synthetic Amino-Terminating Glycans-Implications in Expanding Glycan Microarray Diversity and Platform Comparison. *Glycobiology* (2021), doi:10.1093/glycob/cwab037.
28. E. N. Kitova, L. Han, D. F. Vinals, P. I. Kitov, R. Derda, J. S. Klassen, Influence of labeling on the glycan affinities and specificities of glycan-binding proteins. A case study involving a C-terminal fragment of human galectin-3. *Glycobiology.* 30, 49–57 (2019).
29. C. Büll, R. Nason, L. Sun, J. V. Coillie, D. M. Sørensen, S. J. Moons, Z. Yang, S. Arbitman, S. M. Fernandes, S. Furukawa, R. McBride, C. M. Nycholat, G. J. Adema, J. C. Paulson, R. L. Schnaar, T. J. Boltje, H. Clausen, Y. Narimatsu, Probing the binding specificities of human Siglecs by cell-based glycan arrays. *Proc National Acad Sci.* 118, e2026102118 (2021).
30. W. S. York, R. Mazumder, R. Ranzinger, N. Edwards, R. Kahsay, K. F. Aoki-Kinoshita, M. P. Campbell, R. D. Cummings, T. Feizi, M. Martin, D. A. Natale, N. H. Packer, R. J. Woods, G. Agarwal, S. Arpinar, S. Bhat, J. Blake, L. J. G. Castro, B. Fochtman, J. Gildersleeve, R. Goldman, X. Holmes, V. Jain, S. Kulkarni, R. Mahadik, A. Mehta, R. Mousavi, S. Nakarakommula, R. Navelkar, N. Pattabiraman, M. J. Pierce, K. Ross, P. Vasudev, J. Vora, T. Williamson, W. Zhang, GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology.* 30, 72–73 (2019).
31. V. Padler-Karavani, X. Song, H. Yu, N. Hurtado-Ziola, S. Huang, S. Muthana, H. A. Chokhawala, J. Cheng, A. Verhagen, M. A. Langereis, R. Kleene, M. Schachner, R. J. de Groot, Y. Lasanajak, H. Matsuda, R. Schwab, X. Chen, D. F. Smith, R. D. Cummings, A. Varki, Cross-comparison of Protein Recognition of Sialic Acid Diversity on Two Novel Sialoglycan Microarrays*. *J Biol Chem.* 287, 22593–22608 (2012).
32. J. S. Temme, C. T. Campbell, J. C. Gildersleeve, Factors contributing to variability of glycan microarray binding profiles. *Faraday Discuss.* 219, 90–111 (2019).
33. L. Wang, R. D. Cummings, D. F. Smith, M. Huflejt, C. T. Campbell, J. C. Gildersleeve, J. Q. Gerlach, M. Kilcoyne, L. Joshi, S. Serna, N.-C. Reichardt, N. P. Pera, R. J. Pieters, W. Eng, L. K. Mahal, Cross-platform comparison of glycan microarray formats. *Glycobiology.* 24, 507–517 (2014).
34. Z. Klamer, P. Hsueh, D. Ayala-Talavera, B. Haab, Deciphering Protein Glycosylation by Computational Integration of On-chip Profiling, Glycan-array Data, and Mass Spectrometry* [S]. *Mol Cell Proteomics.* 18, 28–40 (2019).
35. Z. Klamer, B. Haab, (American Chemical Society, 2020; <https://doi.org/10.1021/bk-2019-1346.ch005>), vol. 1346 of Glycan-Based Cellular Communication: Techniques for Carbohydrate-Protein Interactions, pp. 67–82.

36. W. Chai, M. S. Stoll, C. Galustian, A. M. Lawson, T. Feizi, Neoglycolipid Technology: Deciphering Information Content of Glycome. *Methods Enzymol.* 362, 160–195 (2003).
37. J. C. Manimala, T. A. Roach, Z. Li, J. C. Gildersleeve, High-Throughput Carbohydrate Microarray Analysis of 24 Lectins. *Angew Chem-ger Edit.* 118, 3689–3692 (2006).
38. X. Song, Y. Lasanajak, L. J. Olson, M. Boonen, N. M. Dahms, S. Kornfeld, R. D. Cummings, D. F. Smith, Glycan Microarray Analysis of P-type Lectins Reveals Distinct Phosphomannose Glycan Recognition. *J Biol Chem.* 284, 35201–35214 (2009).
39. M. L. Mickum, N. S. Prasanphanich, X. Song, N. Dorabawila, M. Mandalasi, Y. Lasanajak, A. Luyai, W. E. Secor, P. P. Wilkins, I. V. Die, D. F. Smith, A. K. Nyame, R. D. Cummings, C. A. Rivera-Marrero, Identification of Antigenic Glycans from *Schistosoma mansoni* by Using a Shotgun Egg Glycan Microarray. *Infect Immun.* 84, 1371–1386 (2016).
40. Z. Wu, Y. Liu, L. Li, X.-F. Wan, H. Zhu, Y. Guo, M. Wei, W. Guan, P. G. Wang, Decoding glycan protein interactions by a new class of asymmetric N -glycans. *Org Biomol Chem.* 15, 8946–8951 (2017).
41. S. Purohit, T. Li, W. Guan, X. Song, J. Song, Y. Tian, L. Li, A. Sharma, B. Dun, D. Mysona, S. Ghamande, B. Rungruang, R. D. Cummings, P. G. Wang, J.-X. She, Multiplex glycan bead array for high throughput and high content analyses of glycan binding proteins. *Nat Commun.* 9, 258 (2018).
42. L. Liu, A. R. Prudden, C. J. Capicciotti, G. P. Bosman, J.-Y. Yang, D. G. Chapla, K. W. Moremen, G.-J. Boons, Streamlining the chemoenzymatic synthesis of complex N-glycans by a stop and go strategy. *Nat Chem.* 11, 161–169 (2019).
43. C. Gao, M. S. Hanes, L. A. Byrd-Leotis, M. Wei, N. Jia, R. J. Kardish, T. R. McKittrick, D. A. Steinhauer, R. D. Cummings, Unique Binding Specificities of Proteins toward Isomeric Asparagine-Linked Glycans. *Cell Chem Biol.* 26, 535-547.e4 (2019).
44. M. Yan, Y. Zhu, X. Liu, Y. Lasanajak, J. Xiong, J. Lu, X. Lin, D. Ashline, V. Reinhold, D. F. Smith, X. Song, Next-Generation Glycan Microarray Enabled by DNA-Coded Glycan Library and Next-Generation Sequencing Technology. *Anal Chem.* 91, 9221–9228 (2019).
45. N. Wu, L. M. Silva, Y. Liu, Y. Zhang, C. Gao, F. Zhang, L. Fu, Y. Peng, R. Linhardt, T. Kawasaki, B. Mulloy, W. Chai, T. Feizi, Glycan Markers of Human Stem Cells Assigned with Beam Search Arrays*[S]. *Mol Cell Proteomics.* 18, 1981–2002 (2019).
46. G. Murugesan, V. G. Correia, A. S. Palma, W. Chai, C. Li, T. Feizi, E. Martin, B. Laux, A. Franz, K. Fuchs, B. Weigle, P. R. Crocker, Siglec-15 recognition of sialoglycans on tumor cell lines can occur independently of sialyl Tn antigen expression. *Glycobiology* (2020), doi:10.1093/glycob/cwaa048.
47. C. Chen, S. Wang, M. R. Gadi, H. Zhu, F. Liu, C.-C. Liu, L. Li, F. Wang, P. Ling, H. Cao, Enzymatic modular synthesis and microarray assay of poly- N -acetyllactosamine derivatives. *Chem Commun.* 56, 7549–7552 (2020).
48. S. Wang, C. Chen, M. R. Gadi, V. Saikam, D. Liu, H. Zhu, R. Bollag, K. Liu, X. Chen, F. Wang, P. G. Wang, P. Ling, W. Guan, L. Li, Chemoenzymatic modular assembly of O-GalNAc glycans for functional glycomics. *Nat Commun.* 12, 3573 (2021).

49. C. Gao, K. Stavenhagen, B. Eckmair, T. R. McKittrick, A. Y. Mehta, Y. Matsumoto, A. M. McQuillan, M. S. Hanes, D. Eris, K. J. Baker, N. Jia, M. Wei, J. Heimbürg-Molinaro, B. Ernst, R. D. Cummings, Differential recognition of oligomannose isomers by glycan-binding proteins involved in innate and adaptive immunity. Sci Adv. 7, eabf6834 (2021).

Acknowledgments

We thank Dr. Geert-Jan Boons (Heparan Sulfate Array), Dr. Richard Cummings and Akul Mehta (CFG Array), Dr. Ratmir Derda (LiGA), Dr. Fabian Pfrenge (Plant Cell Wall Array), and Dr. Xuezheng Song (NGGM/NGGM-TM) for sharing data in tabular form and/or providing additional experimental information or advice as it relates to their specific arrays.

Funding

National Institute of General Medical Sciences (R44GM131430 and R42GM112750); National Cancer Institute (Early Detection Research Network, U01CA152653; Alliance of Glycobiologists for Cancer Research, U01CA226158)

Author contributions

Conceptualization: ZLK, BBH
Data Curation: ZLK
Software: ZLK
Investigation: CH, JB, JK
Visualization: ZLK, BBH
Supervision: BBH, JZ
Writing—original draft: BBH, ZLK
Writing—review & editing: BBH, ZLK

Competing interests:

JZ is founder and CEO of Z Biotech. The remaining authors declare that they have no competing interests.

Data and materials availability

The models reported here are available through the CarboGrove website. MotifFinder is available as a standalone tool or through a webserver. The webserver features a simplified interface while running the same algorithm as the standalone program. CarboGrove is licensed under the CC BY-SA 4.0 license. Raw glycan array data for published datasets are available from their respective publications. Unpublished glycan-array data are available in the supplementary materials.

Figures and Tables

Table 1. Top 10 glycan-binding proteins for each search motif.

Motif	Rank	Lectin	Name	Canonical Motif	Source	Array
N-Glycan Core a6 Fucose	1	BTL	Bryothamnion Triquetrum Lectin		Investigator	CFG
	2	HML	Hypnea Musciformis Lectin		Investigator	CFG
	3	PSA	Pisum Sativum Agglutinin	Core Fucose	EY Labs	CFG
	4	PSA	Pisum Sativum Agglutinin	Core Fucose	Vector	CFG
	5	rBTL	Bryothamnion Triquetrum Lectin		Investigator	CFG
	6	PSA	Pisum Sativum Agglutinin	Core Fucose	Vector	ASNLinked
	7	LCA	Lens Culinaris Agglutinin	Core Fucose	Vector	ASNLinked
	8	LCA	Lens Culinaris Agglutinin	Core Fucose	Vector	CFG
	9	LCA	Lens Culinaris Agglutinin	Core Fucose	Vector	CFG,NCFG
	10	PSA	Pisum Sativum Agglutinin	Core Fucose	Vector	ZBiotech
Lewis X	1	Anti-LewX	Clone 28 Anti-Lewis X Antibody	Lewis X	Investigator	CFG
	2	LTL	Lotus Tetragonolobus Lectin	Fucose	Vector	PolyLacNAc
	3	CCL2	Coprinopsis Cinerea Lectin 2	Fucose alpha1,3	Investigator	CFG
	4	anti-CD15	anti-CD15	Lewis X	Investigator	PolyLacNAc
	5	AAL	Aleuria Aurantia Lectin	Fucose	Vector	PolyLacNAc
	6	CCL2	Coprinopsis Cinerea Lectin 2	Fucose alpha1,3	Investigator	CFG
	7	AAL	Aleuria Aurantia Lectin	Fucose	Vector	FeiziCLL
	8	LTL	Lotus Tetragonolobus Lectin	Fucose	Vector	Zbiotech
	9	AAL	Aleuria Aurantia Lectin	Fucose	Vector	CEMAOGalNAc
	10	AAL	Aleuria Aurantia Lectin	Fucose	Vector	AsymN
Blood Group B (type 2)	1	EEA	Euonymus Europaeus Agglutinin	Blood Group B	EY Labs	CFG
	2	EEA	Euonymus Europaeus Agglutinin	Blood Group B	Vector	GlycanBead
	3	CGL2	Coprinopsis Cinerea Galectin 2	Fucose alpha1,2	Investigator	CFG
	4	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	CFG
	5	MOA	Marasmius Oreades Agglutinin	Alpha-galactose	EY Labs	CFG
	6	PTA	Psophocarpus Tetragonolobus Agglutinin	Blood Groups	EY Labs	CFG
	7	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	CFG,NCFG
	8	EEA	Euonymus Europaeus Agglutinin	Blood Group B	Vector	CFG
	9	PA-IL	Pseudomonas Aeruginosa Lectin 1	Galactose	Sigma	CFG
	10	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	Zbiotech

Fig. 1. Diversity of glycan-array technology and content. (A) Several technologies in addition to the planar array are now used to probe glycan arrays. The arrays differ in their modes of glycan or protein presentation and in methods of quantification. (B) The sets of glycans contained in the arrays represent diverse types of structures and organisms. (C) The rate of development of new arrays has increased since 2016, punctuated by significant advances in glycan-synthesis technology (red text).

Fig. 2. Standardized analysis and output. (A) MotifFinder analyzes glycan-array data from multiple incubation concentrations, where available, of a given glycan-binding protein. (B) The program identifies the family of motifs the represents the specificity of the protein, organized into primary motifs and the subtrees of fine-specificity motifs. (C) Among the several visualizations in the output are tabular descriptions and binding curves for each motif.

Fig. 3. Breadth of representation of array types of glycan-binding specificities. (A) The collection covers a wide variety of glycan families, array providers, and technologies. **(B)** Motif coverage across the arrays. Nearly all motifs are represented on at least 1 array. **(C)** Motif coverage of the glycan-binding proteins.

Fig. 4. Comparison across multiple arrays of results for SNA. The arrays are ordered from top left to lower right by a reliability score (color-coded based on the scale bar). The relative-binding score of each motif is given next to the ID and graphical representation of the motif.

Fig. 5. Searches for glycan-binding proteins that are specific for selected motifs. (A) The motif scores for 118 motifs (listed in the columns) and 785 glycan-binding proteins were hierarchically clustered. The search motifs and top hits are labeled. **(B)** For the top four glycan-binding proteins from each search, the proteins' top motifs and their relative binding scores are indicated.

Fig. 6. Searches for glycan-binding proteins that are specific for selected motifs. (A) Using models downloaded from CarboGrove that bind any of the comparison motifs, we predicted binding to a series of glycans generated in MotifFinder that contain the motifs, as well as negative-control glycans. **(B)** In the reduced set of 4 lectins (GSL-II, HAA, VVL, and PHA-E), the average binding to the glycans in each group show a different pattern for each group. The differences in top motifs for the lectins correspond to the differences in binding patterns.

ASN-Linked

Gildersleeve

Reichardt

ZBioCatchAll

CFG

CFG - Continued

Microbial

Huflejt

- Monosaccharides
- Glucose
 - Mannose
 - Galactose
 - Neu5Ac
 - GlcNAc
 - GalNAc
 - Fucose
 - Sialic Acid
- Other symbols
- ~SKIP~
 - R-
 - Any # of monosaccharides
 - Hydrogen or any monosaccharide

A Array Technology

Traditional Array

Neoglycoprotein

Glycan Bead

NGGM

CUPRA

LiGA

B Array Content

Man 6-Phosphate

Human Milk Oligosaccharides

Plant Cell Wall Array

Microbial Array

Asymmetric N-Glycan Array

PolyLacNAc

Heparan Sulfate

Chemoenzymatic O-GalNAc

Oligomannose

C Glycan Array Timeline

^a Data Unavailable

^b Array Partially Unsupported

^c Array Unsupported

MotifFinder
Motif Mining

Standardized
Result
Presentation

C

ID	Structure	Relative Binding
A		1.00
B		0.21
C		0.28

Result Quality

ASN-Linked

ZBiotech CatchAll

Gildersleeve

NGGM-TM

AsymmetricN

CFG

ZBiotech
General100

Microbial

HMO

Reichard

Huflejt

NGGM

LiGA

CEMA O-GalNAc

CUPRA

GlycanBead

Reliability Score

Monosaccharides			Other symbols	
● Glucose	■ GlcNAc	◆ Neu5Ac	~SKIP~	Any # of monosaccharides
● Mannose	■ GalNAc	◆ Sialic Acid	R-	Hydrogen or any monosaccharide
● Galactose	▲ Fucose