Combined Analysis of Multiple Glycan-Array Datasets: New Explorations of Protein-Glycan Interactions

Zachary Klamer and Brian Haab

Van Andel Institute, 333 Bostwick NE, Grand Rapids, MI 49503

Supporting Methods

- Curve Fitting Using Results from Multiple Lectin Concentrations
- Mapping Results onto Comparable Scales
- Predicting Lectin Binding to Glycans

Supporting Tables

- Supporting Tables 1. Comparison of MotifFinder and CCARL
- Supporting Tables 2. Array Sources (separate Excel file)

Supporting Figures

- Supporting Figure 1. Identification of Complex, Fine Specificities
- Supporting Figure 2. Comparison of Alternative Array Results to CFG Results
- Supporting Figure 3. MotifFinder Analysis Applied to Glycan-Array Platforms
- Supporting Figure 4. Comparing Model Generalization for Single-Source and Mixed-Source Models
- Supporting Figure 5. Complete Motifs for Comparison of Analysis Tools for the Lectin ECL

Supporting Data

- The MotifFinder output files are provided separately in a single, compressed folder

Supporting Methods

Curve Fitting Using Results from Multiple Lectin Concentrations

Relating data across concentrations can provide valuable insights into the binding to particular motifs but is reliable only for well-controlled datasets collected under consistent conditions. Given such datasets, MotifFinder fits logistic curves to motif binding across arrays, using N-1 parameters (where N is the number of concentrations) up to a 4-parameter logistic regression,

$$A + \frac{-A}{\left(1 + \left(\frac{x}{Kd}\right)^{B}\right)^{C}}$$

where A is the asymptote, B is the hill slope, and C is the asymmetry parameter.

Mapping Results onto Comparable Scales

Integrating data between datasets allows the comparison of glycans that are unique to either array. The lectin concentration must be similar between arrays for the comparison to be valid. But even with this requirement, the scales of quantification can be greatly different, making it necessary to rescale the data. The process of cross-platform normalization is a challenging bioinformatics problem for which many different solutions have been proposed. MotifFinder adopts an approach similar to that of the XPN method[1], with accommodations made for the differences between glycan arrays in content and format.

Like the XPN method, our approach uses a maximum-likelihood estimator to solve for the rescale and shift parameters which maximize the likelihood of obtaining the observed data for one array, given the binding found on the other array. (“Likelihood” is the probability that the observed values in the comparison array could arise from normal distributions derived from the first array.) MotifFinder is different in that it uses average motif binding to estimate binding instead of average cluster binding. Furthermore, MotifFinder allows for motifs to bind in only one or the other array, using weights to reduce the influence of motifs which have low binding in just one array. This step allows motifs to not bind in one array due to differences between the arrays, without throwing off the mapping procedure.

Predicting Lectin Binding to Glycans

MotifFinder includes functionality for applying the binding-specificity model of a lectin to independent glycans in order to predict the level of lectin binding to those glycans. The first step is to identify which of the motifs in the model, if any, each glycan has. In cases where glycans have more than one motif, a unique motif assignment is achieved by following the hierarchical ordering of the regression tree. Next, each glycan is assigned a predicted binding corresponding to the relative binding of the motif at the lectin concentration of the new dataset. For concentrations not found in the training dataset, relative binding is interpolated using either a log-linear interpolation or (when available) the fitted motif logistic curves. In cases where observed glycan binding data is available (as in cross validation) the predicted binding is mapped onto the observed binding scale by solving for scale and shift parameters while optimizing R^2 of the fit. Details of this algorithm have been published previously[2].

References

1. Shabalin, A. A.; Tjelmeland, H.; Fan, C.; Perou, C. M.; Nobel, A. B. Merging Two Gene-Expression Studies Via Cross-Platform Normalization *Bioinformatics* **2008**, 9, 1154-1160.

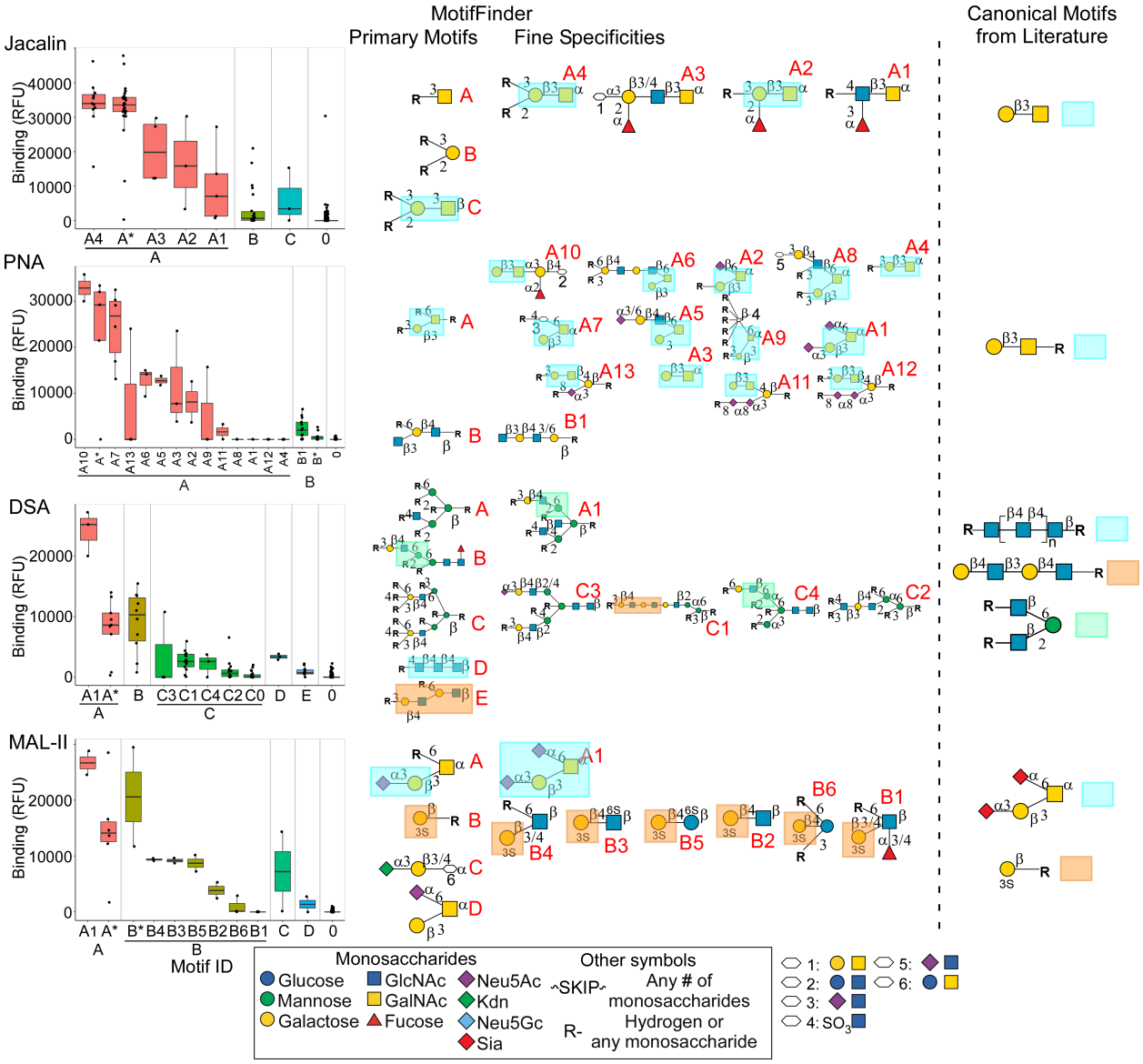
2. Klamer, Z.; Haab, B. Automated Identification of Lectin Fine Specificities from Glycan-Array Data; Glycan-Based Cellular Communication: Techniques for Carbohydrate-Protein Interactions *American Chemical Society* **2020**, 67-82.

Supporting Tables

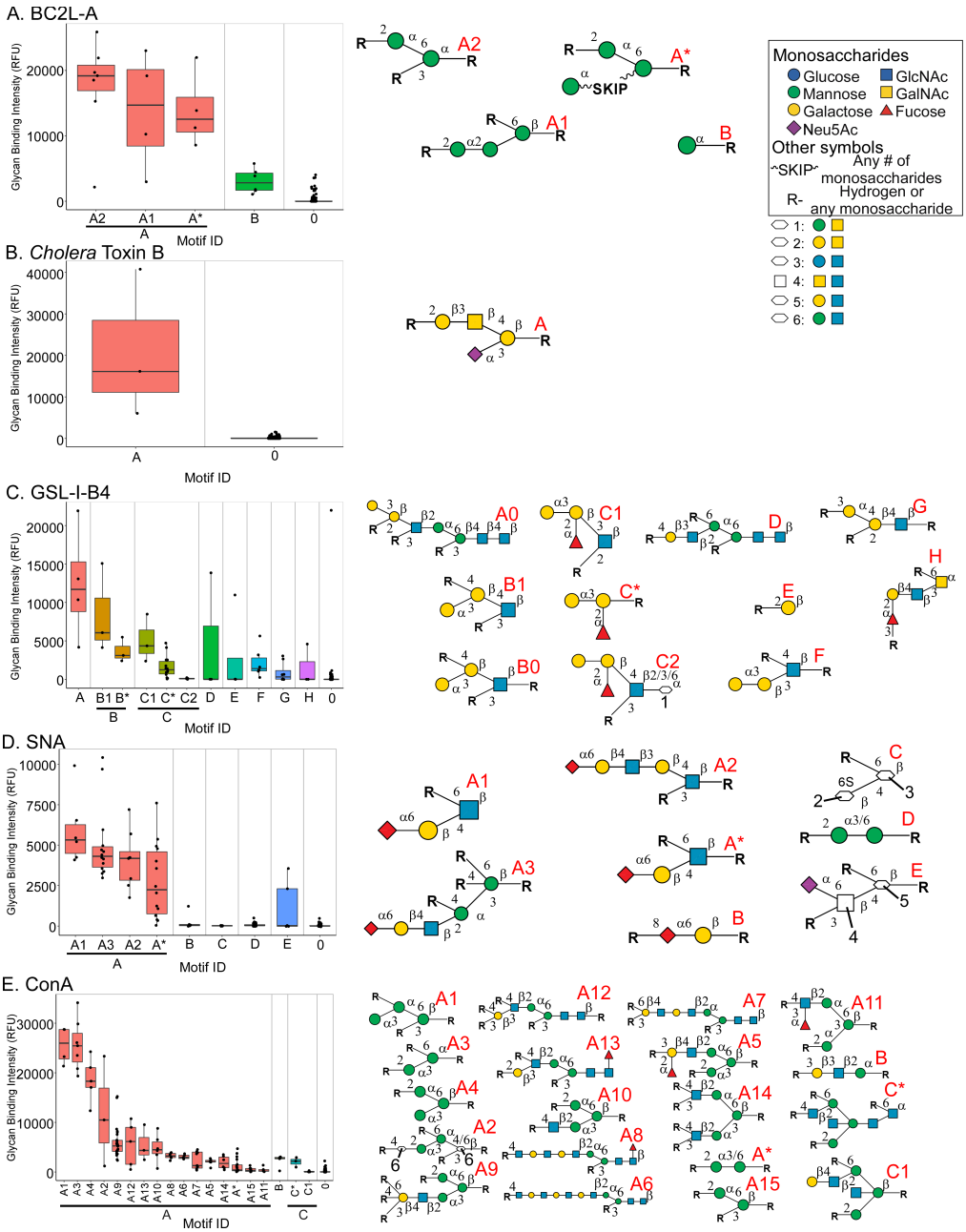
Supporting Table 1. Comparison of MotifFinder and CCARL. Concentrations in bold indicate the concentration used for single concentration analysis (CCARL and MotifFinder Single Concentration) and the concentration used for testing all models. R^2 and RMSE values in bold indicate the best value for the given lectin.

				MotifFinder			
		CCARL		Single Concentration		All Concentrations	
Lectin	Concentration	R^2 (SD)	RMSE (SD)	R^2 (SD)	RMSE (SD)	R^2 (SD)	RMSE (SD)
ABA	1, 10, **100**	0.05 (0.09)	2,924 (1,285)	0.14 (0.09)	2,735 (1,515)	**0.17** (0.10)	**2,624** (1,459)
ConA	1, 5, **10**, 50	0.28 (0.05)	5,880 (179)	**0.48** (0.10)	**4,290** (1,083)	0.46 (0.08)	4,429 (674)
DBA	0.1, 1, 10, **100**	0.20 (0.05)	2,154 (1,225)	0.15 (0.16)	2,240 (1,321)	**0.43** (0.43)	**1,401** (1,496)
DC-Sign	**200**	0.07 (0.08)	**636** (420)	**0.08** (0.10)	640 (451)	N/A	N/A
DSL	0.1, 1, **10**, 100	0.21 (0.03)	2,308 (575)	0.26 (0.17)	2,110 (481)	**0.34** (0.13)	**1,935** (656)
ECL	0.01, 0.1, 0.5, 1, **5**, 50	0.27 (0.01)	1,250 (354)	**0.40** (0.09)	**1,007** (263)	0.37 (0.05)	1,068 (282)
GSL-I-B4	0.5, 1, **10**, 100	0.29 (0.08)	1,569 (929)	0.08 (0.09)	1,899 (989)	0.15 (0.11)	1,802 (1,040)
H1N1	**200**	0.04 (0.16)	526 (337)	**0.14** (0.11)	**496** (361)	N/A	N/A
H3N8	**200**	0.21 (0.10)	**59** (25)	**0.23** (0.09)	62 (34)	N/A	N/A
Jacalin	0.1, **1**, 10, 100	0.12 (0.02)	6,281 (1,014)	0.45 (0.13)	3,812 (597)	**0.50** (0.12)	**3,524** (675)
LCA	0.1, 1, **10**, 100	0.28 (0.03)	5,995 (763)	0.44 (0.30)	4,716 (2,745)	**0.51** (0.18)	**3,998** (1,450)
MAL-I	0.1, 1, **10**, 100	0.24 (0.03)	1,246 (206)	**0.38** (0.15)	**1,014** (285)	0.36 (0.19)	1,050 (404)
MAL-II	0.1, 1, **10**, 100	0.22 (0.01)	2,460 (1,056)	0.38 (0.23)	2,004 (1,306)	**0.39** (0.24)	**1,960** (1,309)
PHA-E	0.1, 1, **10**, 100	0.30 (0.03)	5,487 (1,099)	**0.34** (0.13)	**5,087** (970)	0.30 (0.08)	5,449 (1,029)
PHA-L	0.1, 1, **10**, 100	0.22 (0.03)	2,461 (716)	**0.56** (0.21)	**1,502** (1,007)	0.53 (0.14)	1,573 (859)
PNA	0.1, 1, **10**, 100	0.21 (0.02)	3,748 (1,951)	**0.45** (0.15)	**2,501** (1,283)	0.40 (0.20)	2,992 (2,409)
PSA	0.1, 1, **10**, 100	0.29 (0.02)	3,869 (783)	0.44 (0.16)	3,059 (1,235)	**0.47** (0.13)	**2,890** (896)
RCA-I	0.1, 1, **10**, 100	0.19 (0.05)	10,327 (438)	0.47 (0.08)	6,748 (869)	**0.49** (0.08)	**6,554** (972)
SBA	0.1, 1, **10**, 100	0.23 (0.02)	4,049 (920)	**0.47** (0.22)	**2,916** (1,709)	0.46 (0.19)	2,958 (1,618)
SNA	0.005, 0.01, 0.1, 0.5, 1, 5, **10**, 50	0.24 (0.07)	7,677 (617)	**0.63** (0.13)	**3,659** (958)	0.63 (0.12)	3,733 (1,016)
UEA-I	0.1, **1**, 100	0.27 (0.03)	4,449 (1,523)	0.32 (0.15)	4,138 (1,997)	**0.40** (0.17)	**3,751** (2,049)
WGA	0.01, 0.1, **1**, 10, 100	0.11 (0.04)	8,853 (1,633)	0.17 (0.06)	8,319 (1,800)	**0.18** (0.07)	**8,211** (1,937)
TOTAL		0.21 (0.08)	3,828 (2,793)	0.34 (0.16)	2,952 (2,049)	**0.40** (0.13)	**3,258** (1,898)

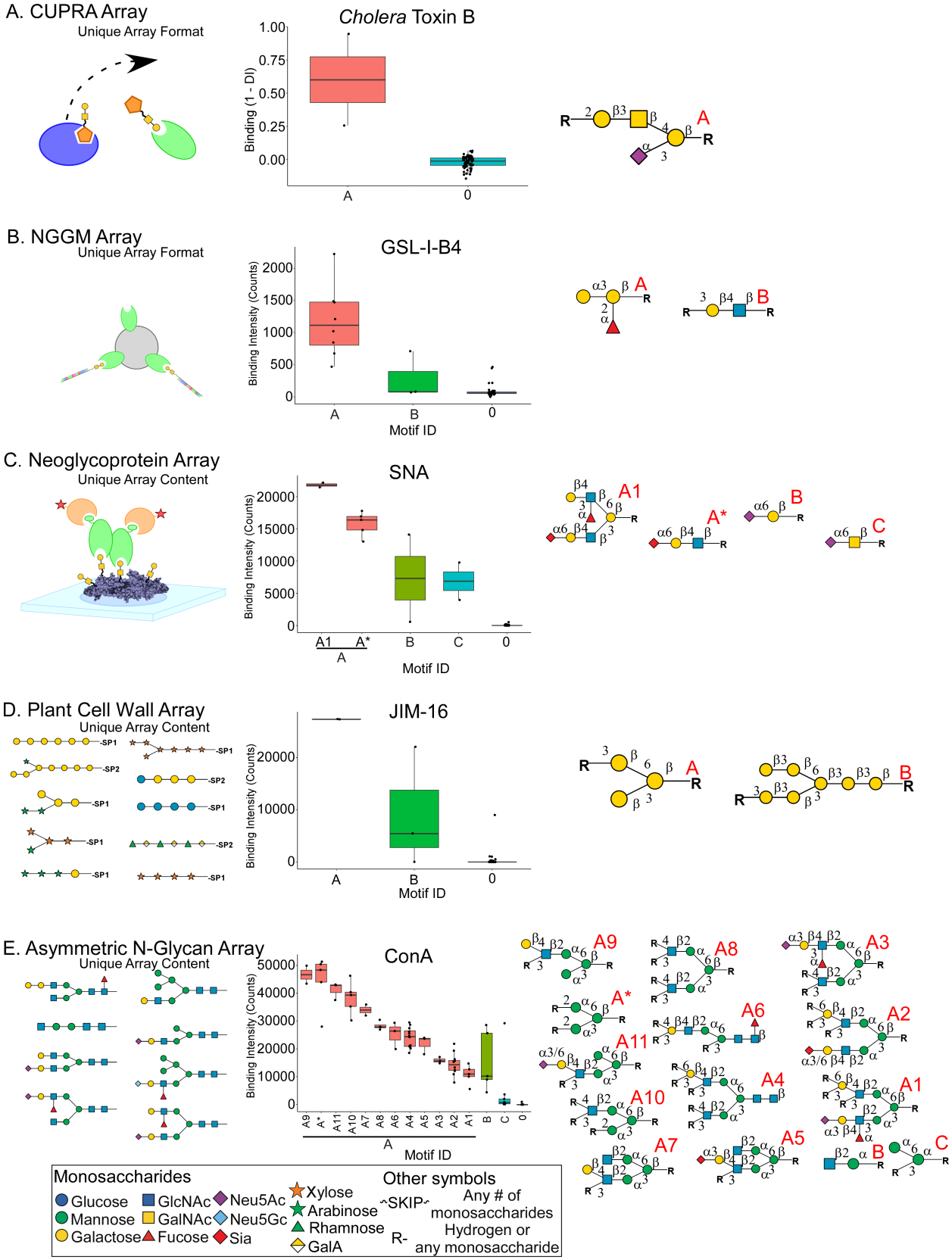
Supporting Figures



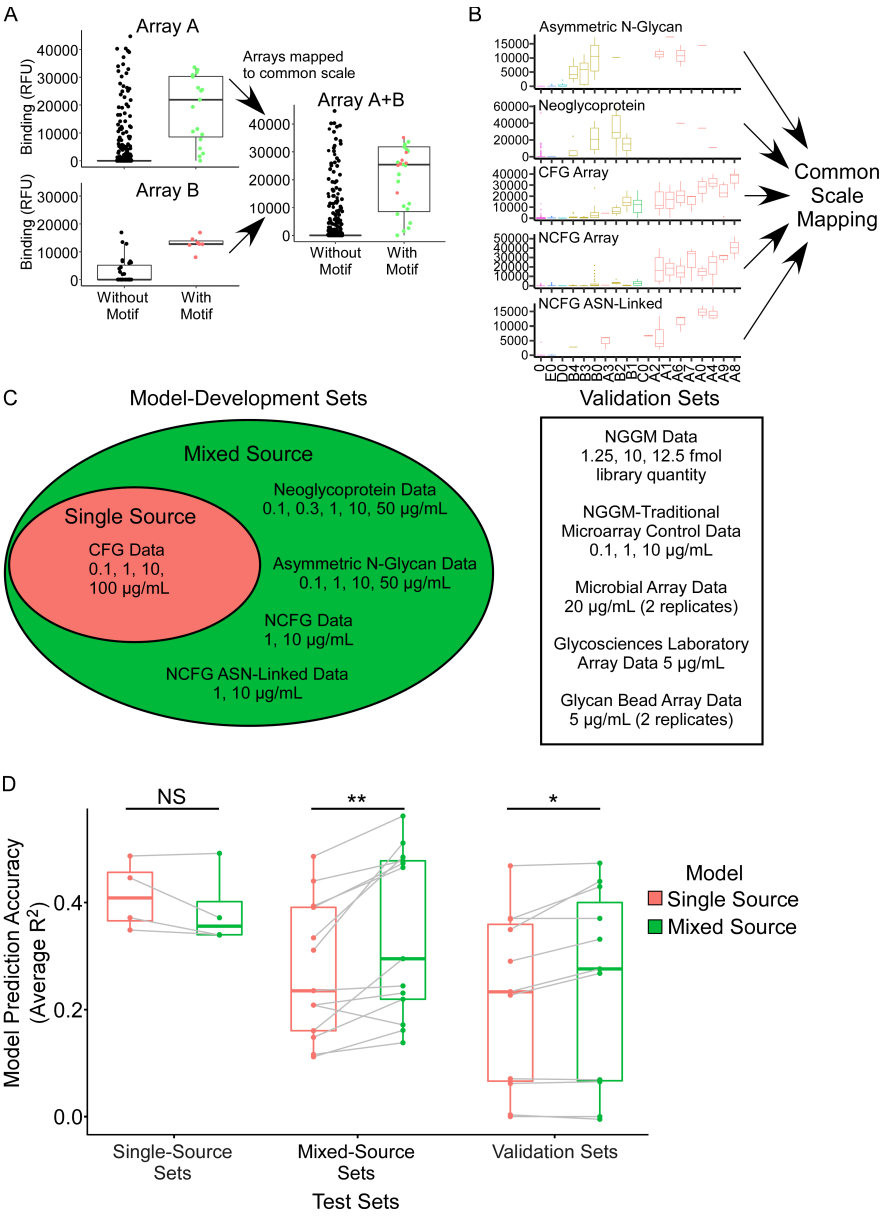
Supporting Figure 1. Identification of Complex, Fine Specificities. The binding intensities to the glycans (graphs on the left) are grouped according to their primary motifs or fine specificities. Each point indicates a unique glycan. In the graphical representations of the motifs (right), the colored boxes indicate the locations of the canonical motifs in the MotifFinder motifs.



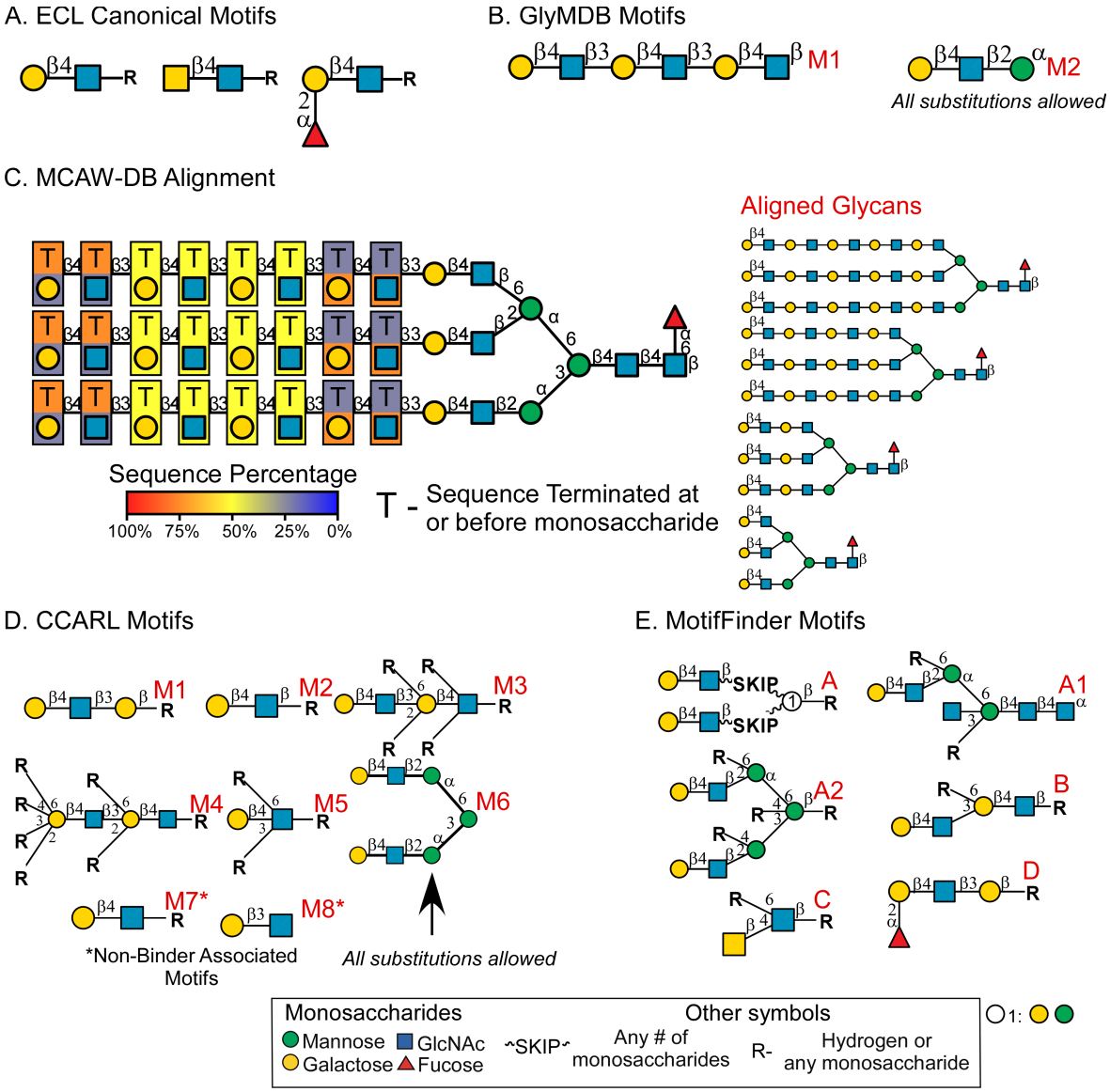
Supporting Figure 2. Comparison of Alternative Array Results to CFG Results. The lectins presented here correspond to the lectins used in the non-CFG arrays (Fig. 2 of the main text and Supplementary Fig. 3). A. BC2L-A binding to the CFG array reveals only high-mannose N-glycan epitopes, in contrast to results from the microbial array. B. *Cholera* toxin B binding to the CFG array shows the same motif as found on the CUPRA array (Supplementary Fig. 3A). C. GSL-I-B4 analysis on the CFG array shows a greater range of fine specificities, particularly among alpha- galactose structures, than the analysis on NGGM array (Supplementary Fig. 3B). D. SNA binding is similar to that found on the neoglycoprotein array (Supplementary Fig. 3C) with some additional fine specificities. E. ConA binding to the CFG array shows similarity to the binding to the asymmetric N glycan array (Supplementary Fig. 2D) but lacks definition of asymmetric motifs.


A. CUPRA Array
Unique Array Format
Cholera Toxin B
Binding (1 - DI)
1.00
0.75
0.50
0.25
0.00
A
0
B. NGGM Array
Unique Array Format
GSL-I-B4
Binding Intensity (Counts)
2000
1500
1000
500
0
A
B
0
Motif ID
C. Neoglycoprotein Array
Unique Array Content
SNA
Binding Intensity (Counts)
20000
15000
10000
5000
0
A1
A*
B
C
0
A
Motif ID
D. Plant Cell Wall Array
Unique Array Content
JIM-16
Binding Intensity (Counts)
20000
10000
0
A
B
0
Motif ID
E. Asymmetric N-Glycan Array
Unique Array Content
ConA
Binding Intensity (Counts)
50000
40000
30000
20000
10000
0
A9
A*
A11
A10
A7
A8
A6
A4
A5
A2
A1
B
C
0
A
Motif ID
Monosaccharides
Glucose
Mannose
Galactose
GlcNAc
GalNAc
Fucose
Neu5Ac
Neu5Gc
Sia
Xylose
Arabinose
Rhamnose
GalA
Other symbols
SKIP
Any # of monosaccharides
R-
Hydrogen or any monosaccharide

Supporting Figure 3. MotifFinder Applied to Additional Glycan-Array Platforms. A. The CUPRA array uses a depletion index (DI) to indicate the depletion of a proxy receptor. B. The NGGM (Next Generation Glycan Microarray) uses DNA barcodes for quantification. MotifFinder correctly found alpha-linked Gal as the main motif for GSL-I-B4. C. The MotifFinder analysis of SNA assayed on a neoglycoprotein array correctly identified binding to alpha-6 linked sialic acids in various contexts. D. An array of glycan structures from the cell walls of plants enables probing with anti-cell-wall antibodies such JIM-16. The MotifFinder results match the published results, with additional information about gradations. E. An array of asymmetrically branched N-glycans allows the testing of branch-dependent binding, as demonstrated here for ConA. The fine-specificities of motif A reveal variation within the primary motif of the N-glycan core (motif A*).



Supporting Figure 4. Comparing Model Generalization for Single-Source and Mixed-Source Models. A. MotifFinder maps the various datasets onto a common scale to allow comparisons of glycans not found on both arrays. B. For five datasets collected with RCA-I, the diverse ranges of RFU values were mapped onto a common scale for direct comparison. C. Models were generated using data from only one source or, alternatively, from multiple sources. The datasets used in type of model are indicated by the Venn diagram. The models were applied to predict lectin binding to the glycans in the validation sets listed at right, which were not used in the creation of the models. D. Using 5-fold cross validation, the models that were generated from 4/5 of the glycans were used to predict lectin binding to the remaining 1/5, and the average R^2 between predicted and observed binding was calculated. Each point represents an individual dataset. ** indicates $p < 0.01$; * indicates $p < 0.05$; NS, not significant, paired t-test with unequal variance.



Supporting Figure 5. Complete Motifs for Comparison of Analysis Tools for the Lectin ECL. A. The canonical motifs for ECL binding. B. The motifs obtained using the GlyMDB tool. These motifs are simple subtrees and therefore do not specify substitution intolerance such as a terminal epitope. C. A graphical representation of the MCAW-DB alignment of top glycans. The alignment as presented can be found in the MCAW-DB (https://mcawdb.glycoinfo.org/detail.html?737). The top-binding glycans used for the alignment are given (right). D. The motifs obtained using the CCARL method. For readability, all substitution tolerant carbons in motif M6 are not noted as all carbons are substitution tolerant. E. The motifs obtained using MotifFinder. These motifs include motifs C and D which are similar to canonical motifs not found with other methods.