Correcting for sparsity and non-independence in glycomic data through a systems biology framework

Bokan Bao[1,2,3,+], Benjamin P. Kellman[1,2,3,+], Austin W.T. Chiang[1,2], Austin K. York[1,2], Mahmoud A. Mohammad[4], Morey W. Haymond[4], Lars Bode[1], Nathan E. Lewis[1,2,5,*]

[1] Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

[2] The Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, La Jolla, CA 92093, USA

[3] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA

[4] Department of Pediatrics, Children's Nutrition Research Center, US Department of Agriculture/Agricultural Research Service, Baylor College of Medicine, Houston, Texas 77030, USA

[5] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

[+] These authors contributed equally to this work

[*] Corresponding author:

Name: Nathan E. Lewis

Address: 9500 Gilman Drive MC 0760, La Jolla, CA 92093

Phone: 858-246-1876

E-mail: nlewisres@ucsd.edu

Short running title: Comparative glycoprofiles analysis with glycan substructures

Abstract

Glycans are fundamental cellular building blocks, involved in many organismal functions. Advances in glycomics are elucidating the roles of glycans, but it remains challenging to properly analyze large glycomics datasets, since the data are sparse (each sample often has a minority of measured glycans) and detected glycans are non-independent, sharing many intermediate biosynthetic steps with other glycans. We address these challenges with GlyCompare, a glycomic data analysis approach that leverages shared biosynthetic pathway intermediates to correct for sparsity and non-independence in glycomics. Specifically, quantities of measured glycan are propagated to intermediate glycan substructures, which enable direct comparison of different glycoprofiles. Using this, we studied diverse N-glycan profiles from glycoengineered erythropoietin. We obtained biologically meaningful clustering of mutant cells and identified knockout-specific effects of fucosyltransferase mutants on tetra-antennary structures. We further analyzed human milk oligosaccharide profiles and identified novel impacts that the mother's secretor-status has on fucosylation and sialylation. Our GlyCompare substructure-oriented approach will enable researchers to take full advantage of the growing power and size of glycomics data.

Introduction

Glycosylation is a highly abundant and complex post-translational modification, decorating between one-fifth and one-half of eukaryotic proteins[1,2]. These diverse carbohydrates account for 12-25% of dry mass, and have important functional and pathological roles[3,4]. Despite their essentiality, glycans are difficult to study; thus, our understanding of their impact is limited. The complex structures of glycans arise from a non-template driven synthesis through a biosynthetic network. A simple change of a single intermediate glycan or glycosyltransferase will have cascading impacts on the final glycans obtained[5,6]. Unfortunately, glycoprofiling and glycomic data lack the necessary systems perspective to easily decode this potential for change. It is important to understand the network behind the glycoprofiles so that we can better understand the regulation of this process.

New tools that aid in the acquisition and aggregation of glycoprofiles are emerging, making large-scale comparisons of glycoprofiles possible. Advances in mass spectrometry now enable the rapid generation of large volumes of glycoprofiles with detailed glycan composition[7–10], exposing the complex and heterogeneous glycosylation patterns on lipids and proteins[11,12]. Large glycoprofile datasets and supporting databases are also emerging, including GlyTouCan[13], UnicarbDB[14], Glygen and UniCarbKB[15].

These new technologies and databases provide opportunities to examine global trends in glycan function and their association with disease. However, the rapid and accurate comparison of glycoprofiles can be challenging with the size, sparsity and heterogeneity of such datasets. Indeed in any one glycoprofiles, only a few glycans may be detected among the thousands of possible glycans. Thus, even in the same dataset, few glycans, if any, overlap between samples if there is a major perturbation to glycosylation. However, these non-overlapping glycans may only

differ in their synthesis by as few as one enzymatic step. Thus, it can be difficult to know which glycans to compare. Furthermore, since glycans often share substantion portions of their biosynthetic pathways with each other, statistical methods that assume independence (e.g., t-tests, ANOVA, etc) are inappropriate for glycomics. Here we address these challenges by proposing that glycan substructures, or intermediates, are useful functional units to enable meaningful glycoprofile comparisons, since each substructure can capture one step in the complex process of glycan synthesis. Thus, using substructures for comparison, we account for the shared dependencies of all glycans.

Previous work has investigated the similarity across glycans using glycan motifs, such as glycan fingerprinting to describe glycan diversity in glycan databases[16], alignment of glycan structures[17], identification of glycan epitopes in glycoprofiles[18], graph-based deconvolution of LC-MS to clarify glycan abundance[19], or comparison of glycans in glycoprofiles leveraging simple structures[20]. These tools use information on glycan composition or epitopes. However, accounting for the shared biosynthetic steps in the various glycans in glycoprofiles, provides complete biosynthetic context to all glycan epitopes. That context includes connecting all glycans to the enzymes involved in their synthesis, the order of the enzyme reactions, and information on competition for glycan substrates. Thus, a generalized substructure approach can facilitate the study of large numbers of glycoprofiles and connect them to the shared mechanisms involved in making each glycan.

Here we present GlyCompare, a method enabling the rapid and scalable analysis and comparison of any number of glycoprofiles while accounting for the biosynthetic similarities of each glycan. This will facilitate the discovery of mechanisms underlying the changes among glycoprofiles. We demonstrate the functionalities and performance of this approach with both

protein-conjugated and unconjugated glycomics, using recombinant erythropoietin (EPO) N-glycosylation and human milk oligosaccharides (HMOs). Specifically, we analyzed sixteen MALDI-TOF glycoprofiles of EPO, where each glycoprofile is of EPO produced in a different glycoengineered CHO cell line[9,11]. We also analyzed forty-eight HMO glycoprofiles from six mothers[21]. By analyzing these glycoprofiles with GlyCompare, we quantify the abundance of important substructures, cluster the glycoprofiles of mutant cell lines, connect genotypes to unexpected changes in glycoprofiles, and associate a phenotype of interest with substructure abundance and synthesis network dynamics. The analysis of the EPO and HMO datasets demonstrate that our novel framework presents a convenient and automated approach to elucidate novel insights into complex patterns in glycobiology.

Results

GlyCompare decomposes glycoprofiles to facilitate glycoprofile comparison

Glycoprofiles can be decomposed into glycan substructure abundances. The resulting substructure profile has richer information than whole glycans and enables more precise comparison across conditions. However, the comparison does not need all substructures. Glycan biosynthesis involves long pathways with multiple intermediates and much pathway redundancy. Thus, pathways can be collapsed to obtain a subset of substructures while preserving the information of all glycans in the dataset. We call this minimal set of substructures “glyco-motifs”. The GlyCompare workflow consists of several steps wherein glycoprofiles are annotated and decomposed, glyco-motifs are prioritized, and each glyco-motif is quantified for subsequent condition-specific comparisons (**Fig. 1**). The specific workflow is described as follows.

First, to characterize each glycoprofile by substructures, all relevant substructures are defined and quantified. Thus, a complete set of glycan substructures is obtained for all glycans in all glycoprofiles being analyzed. For each glycoprofile, the abundance of each substructure is calculated by summing the abundance of all glycans associated with the substructure. This results in a substructure profile, which stores abundances for all glycan substructures (**Fig. 1a,b**) in given glycoprofile.

Second, to identify the most informative substructures (glyco-motifs), substructures are prioritized using the substructure network. The substructure network is built by connecting all substructures into their biosynthetic pathways (**Fig. 1c**). To improve computational efficiency, only substructures associated with glycans being analyzed are included in the substructure network. Starting from the monosaccharides, each level of the network represents another biosynthetic step, with one more monosaccharide than the previous level. The edges in the network are the enzymatic reactions adding each monosaccharide. These edges are weighted by the correlation between the abundances of the substrate and product substructures across all samples. Redundant substructures can be easily identified since their parent-child substructure abundances will be perfectly correlated. Substructure network reduction proceeds by collapsing links with a perfect correlation between substrate and product substructures, and only retaining the product substructure. The remaining substructures are termed glyco-motifs; they completely describe the variance at the substructure level. The abundances of all glyco-motifs are then represented with a glyco-motif profile, the minimal subset of meaningful substructure abundances (**Fig. 1d**).

For larger datasets, summarizing the glyco-motifs becomes necessary. Since the glyco-motifs are also clustered when clustering the glycoprofiles (**Fig. 1d**), representative substructures are

generated by summarizing the common structure in each cluster (**Fig. 1e**). The representative substructure describes the glycan features that vary the most across samples. To extract the common structure features, the average weight of each monosaccharide is calculated and the monosaccharides with a weight larger than 51% are preserved, which illustrates the predominant structure in the cluster. This allows one to quickly evaluate the distinguishing glycan features that vary across samples in any given dataset.

The workflow we described here successfully connects all glycoprofiles in a data set by their shared intermediate substructures, thus allowing robust analysis of the differences across glycomics samples and the evaluation of the associated genetic bases.

GlyCompare accurately clusters glycoengineered EPO samples

In examining the EPO glycosylation data[9], we found the glycoprofiles failed to form robust clusters that can be easily interpreted. Furthermore, glycoprofiles with low genetic similarity clustered together (**Fig. 2a and Supplementary Fig. 1**, 2). This inconsistency and poor clustering stems from the inherent sparseness of the glycoprofile datatype, i.e., each glycoprofile only has a few glycans, making the matrix of all samples very sparse, unfit for standard clustering approaches, and thus hard to interpret. Particularly problematic is that pairs of glycans differing in a single monosaccharide are treated as two completely different glycans under standard clustering approaches. Thus, we found that clustering is affected more by the presence or absence of a glycan, rather than structural similarity.

GlyCompare addresses these problems by elucidating hidden similarities between glycans after decomposing glycoprofiles to their composite substructures. The 52 glycans were decomposed into their constituent glycan substructures, resulting in a substructure profile that

consists of 613 glycan substructures and a further simplified 120 glyco-motif profile (**Supplementary Fig. 3**). The clustering automatically distinguished the samples based on the structural patterns, and clustered the glyco-motif profiles into groups more consistently associated with the extent of changes in the profile than the raw glycan-based clusters (**Fig. 2b and Supplementary Fig. 4, 5**).

The sixteen glycoprofiles clustered into three groups with a few outliers (**Fig. 2b**), and the 120 glyco-motifs clustered into twenty-four groups, each summarized by representative substructures (**Fig. 2c and Supplementary Fig. 5**). The clusters of glycoprofiles are consistent with the genetic similarities among the host cells. Specifically, the major substructure patterns cluster individual samples into groups of 'wild-type (WT)-like', 'mild', 'medium' and 'outlier' categories. The WT-like category contains one group, WT and B4galt1/2/3/4/ knockouts, which contains most of the substructures seen in WT cells. The mild group includes the Mgat4b/4a, Mgat4b, and Mgat5 knockouts, which each lose the tetra-antennary structure, and an individual St3gal4/6 knockout that loses the terminal sialylation. The medium category is a group that contains knockouts of St3gal4/6 and Mgat4a/4b/5, knockouts of Mgat4a/4b/5 and B3gnt2, knockouts of Mgat4a/4a/5 with a knock-in of human ST6GAL1, and knockouts of Mgat4a/4b/5 and St3gal4/6, which lose the tri-antennary structure. The 'outlier' category includes three individual glycoprofiles with knockouts for Fut8, Mgat2, and Mgat1, each of which generate many glycans not detected in the WT nor other three groups ('WT-like', 'mild' and 'medium'). While some glyco-motif clusters can be seen in the glycoprofile clusters, there are important differences, and the glyco-motif clusters provide more information and improved cluster stability (**Supplementary Fig. 2**). The results demonstrated the excellent performance of our GlyCompare in assessing the structural similarity between different glycoprofiles.

GlyCompare identifies substructure changes across glycoprofiles

GlyCompare helps to more robustly group samples by accounting for the biosynthetic and structural similarities of glycans. Further analysis of the representative structures provides detailed insights into which structural features vary the most across samples. To accomplish this, we standardized--using z-score normalization--the representative structure abundances. We then identified significant changes in representative substructure abundances between mutant cells and WT (**Fig. 2d**). This highlights the specific structural features of glycans that are impacted when glycoengineering recombinant EPO.

As expected, in the Mgat1 knockout glycoprofile, only high mannose N-glycans are seen. Also, in the Mgat2 knockout, the glycan substructure of bi-antennary on one mannose linkage significantly increases, and the unique structure of bi-antennary LacNac elongated in the N-glycans emerges in the St3gal4/6 and Mgat4a/4b/5 knockouts. Along with expected changes in α-1,6 fucosylation in the Fut8 knockout glycoprofile, we also observed an increase in the tetra-antennary poly-LacNac elongated N-glycan without fucose, which has not been previously reported (**Fig. 3a**). In the St3gal4/6 knockout (**Fig. 3a**), we observed the relative abundance of structures with sialylation has significantly decreased, where the tetra-antennary and triantennary poly-LacNAc elongated N-glycan substructure without sialylation significantly increased (Rep13: $p<0.01$, Rep20: $p<0.001$). Finally, in the Mgat4b, Mgat4a/4b and Mgat5 knockout glycoprofiles (**Fig. 3b**), tetra-antennary substructures (Rep16-Rep22) decreased significantly ($p<0.01$), while triantennary substructures with GlcNac elongation (Rep13-Rep14) increased significantly ($p<0.01$). However, the poly-LacNac elongation structure decreased significantly for the Mgat5 and Mgat4b knockout ($p<0.001$). Interestingly, while both the Mgat4b and Mgat5

knockouts significantly decreased in the triantennary poly-LacNac elongated N-glycan, the Mgat4b knockout significantly increased in the tri-antennary LacNac elongated N-glycan. Thus, through the use of GlyCompare, we identified the specific glycan features that are impacted not only in individual glycoengineered cell lines, but also features shared by groups of related cell lines.

GlyCompare reveals phenotype-associated substructures and trends invisible at the whole glycan level

Many secreted and measured glycans are also precursors, or substructures, of larger glycans (**Fig. 4a**). Thus, the secreted and observed abundance of one glycan does not equal to the total amount synthesized. GlyCompare can quantify the total abundance of a glycan by combining the glycan abundance with the abundance of its products. To demonstrate this capability of GlyCompare, we analyzed HMO abundance, and examined the impact of secretor status and days postpartum on HMO abundance. We obtained forty-seven HMO glycoprofiles from 6 mothers (1, 2, 3, 4, 7, 14, 28 and 42 days postpartum (DPP)), 4 “secretor” mothers with functioning FUT2 (α-1,2 fucosyltransferase), and 2 “non-secretor” mothers with non-functional FUT2. With the non-independence of HMOs addressed with GlyCompare, we were able to use powerful statistical methods to study trends in HMO synthesis. Specifically, we used Generalized Estimating Equations[22] to build regression models predicting secretor status and DPP from substructure abundance.

We first checked both the glycan-level and substructure-level clustering of the glycoprofile. Samples with same secretor status and days postpartum (DPP) were successfully grouped (**Supplementary Fig. 6**). Further examination of the total amount of substructure synthesized

revealed phenotype-related trends invisible at the level of the raw glycoprofile. For example, the LSTb substructure (X80) increased in secretor mothers (Wald $p = 0.001$) and decreased in non-secretor mothers over time (Wald $p < 2.2x10^{-16}$; **Fig. 4b**). Yet, the same trend was weak or inconsistent for all glycans containing the X80 substructure: LSTb, DSLNT and DSLNH. Like X80, LSTb and DSLNT decreased in abundance over time in non-secretors (LSTb Wald $p=0.0046$, DSLNT Wald $p = 0.0045$); however, unlike X80, neither LSTb nor DSLNT significantly increased in secretor mothers over time (Wald $p>0.15$). Meanwhile, DSLNH, like X80, increased over time in secretor mothers (Wald $p=0.0062$) (**Fig. 4e**). Rather than decreasing like X80, non-secretor DSLNT abundance significantly increased over time (Wald $p=5.8x10^{-9}$) (**Fig. 4d**). The secretor-specific trends in total LSTb are only visible examining the X80 substructure abundance (**Fig. 4c**). Thus, while secretor status is expected to impact HMO fucosylation, GlyCompare reveals associations with non-fucosylated substructures. Viewing substructure abundance as total substructure synthesized provides a new fundamental measure to the study of glycoprofiles, it also creates an opportunity to explore trends in synthesis.

GlyCompare identifies condition-specific synthesis dynamics

We applied GlyCompare to explore changes in HMO synthesis over time. We estimate the fluxes for each biosynthetic reaction by quantifying the ratio of product-reactant pairs (pairs of glycan substructures that differ by one monosaccharide). By comparing the substructure abundance of reactants and products, we can estimate how many reactant substructures are converted to product substructure. With clearly defined glycan biosynthesis pathways, we explore synthesis dynamics through the systematic estimation of reaction flux under various conditions.

We measured the product/reactant ratio for every reaction in the HMO substructure network (**Supplementary Fig. 7**) to estimate flux for each reaction over time and across secretor status. We found several reactions strongly associated with secretor status (**Supplementary Table 1 and Supplementary Fig. 8**), i.e. the ability of the mother to add an α-1,2 fucose to HMOs. As expected, the reaction flux from the LNT substructure (X38) to the LNFPI substructure (X86), was strongly associated with secretor status (Wald $p<2.2x10^{-16}$) (**Fig. 4a**). In secretors, 36.2% of X38 was converted to X86, compared to non-secretors, wherein only 5% of X38 was converted.

Eventhough secretor status is defined by the fucosyltransferase-2 genotype, not all secretor-associated reactions were fucosylation reactions. We further explored the secretor-X80 association using the product/reactant ratio. Specifically, we examined the upstream reaction of LNT (X38) to LSTb (X80) and the downstream reaction--LSTb (X80) to DSLNT (X103) (**Fig. 4b, 5a**). Using the product/reaction ratio of substructure abundances, we found secretor status correlates with the reaction of LSTb to DSLNT but does not impact the conversion of LNT to LSTb; any changes in sialylation is intriguing considering secretor status only directly impacts fucosylation. We measured the upstream reaction of LNT converting to LSTb, using the X80/X38 ratio over time; however, no significant change was observed with respect to secretor status (Wald p=0.44). In the reaction of LSTb to DSLNT, we found a secretor-specific reaction propensity. Specifically, we quantified the association of the X103/X80 ratio with secretor status. The X103/X80 ratio was significantly higher (Wald p=0.018) in secretor mothers (**Fig. 5b)**. The higher ratio suggests that the turnover from LSTb (X80) to DSLNT (X103) is significantly higher in secretor mothers. In the average non-secretor mother, 52.3% of LSTb is converted to DSLNT. Meanwhile in secretors, the average conversion rate is 81.8% (a 63.9% reduction in the efficiency of the LSTb to DSLNT conversion). The LSTb to DSLNT conversion rate appears

higher in secretors while conversion from the LSTb precursor, LNT, appears unchanged. Examining the product/reactant ratio has revealed a phenotype-specific reaction propensity thus providing insight to the condition-specific synthesis.

Discussion

Glycosylation has generally been studied from the whole-glycan perspective using mass spectrometry and other analytical methods. From this perspective, two glycans that differ by only one monosaccharide are completely distinct and are not comparable. Thus, the comparative study of glycoprofiles has been limited to changes between glycans shared by multiple glycoprofiles or small manually curated glycan substructures[16]. Glycompare sheds light on the hidden biosynthetic interdependencies between glycans by integrating the biosynthetic pathways into the comparison. Glycoprofiles are converted to glyco-motif profiles, wherein each substructure abundance represents the cumulative abundance of all glycans containing that substructure. This enumeration and quantification of substructures can be easily scaled up to include many glycoprofiles in large datasets. Additionally, since no prior information is required beyond glycan identities and quantities, the method can even facilitate analysis of glycans with limited characterization. Thus, it brings several advantages and new perspectives to enable the systematic study of glycobiology.

First, the GlyCompare platform computes a glyco-motif profile (i.e., the abundances of the minimal set of glycan substructures) that maintains the information of the original glycoprofiles, while exposing the shared intermediates of measured glycans. These sample-specific glyco-motif profiles more accurately quantify similarities across glycoprofiles. This is made possible since

glycans that share substructures also share many biosynthetic steps. If the glycan biosynthetic network is perturbed, all glycans synthesized will be impacted and the nearest substructures will directly highlight where the change occurred. For example, in EPO glycoprofiles studied here, the tetra-antennary structure is depleted on the Mgat4A/4B/5 knockout group and the downstream sialylated substructure depleted when St3gal4/6 were knocked out. Such structural patterns emerge in GlyCompare since the tool exposes the substructures through clustering, thus identifying common features in glycans measured across diverse samples.

Second, new trends in synthesis network dynamics become visible at the substructure level. For example, in the HMO data set, multiple HMOs are made through a series of steps from LNT to DSLNH (**Fig. 4a**). Only when the substructure abundances are computed are we able to observe the secretor-dependent differences in the LSTb substructure, X80. This is particularly interesting since secretor status is defined by changes in α-1,2 fucosylation, but we see here additional secretor-dependent changes to sialylated structures with no fucose. These are the systemic effects invisible without a systems-level perspective due to the interconnected nature of glycan synthesis; this disparity underlines the power of this method.

Third, the sparse nature of glycobiology datasets and the inherent connections between glycans make glycomic data unfit for many statistical analyses commonly used. However, the translation of glycoprofiles into substructure abundance provides a framework for more powerful and statistically rigorous analysis of glycomic datasets. Single sample perturbations, such as the knockouts in the glycoengineered EPO, can be compared to wild-type; all substructure data can be normalized and then rigorously distinguished from the control using a one sample t-test. Conditions or phenotypes with many samples, such as the secretor status in the HMO dataset, can be compared using regression models, like Generalized Estimating Equations (GEE)[22], to

evaluate the association between the phenotypes and glycosylation. For example, in HMO data, the GEE revealed that the α-1,2 fucose substructure is enriched in secretor status, consistent with the previous studies[23–25]. Additionally, the GEE discovered that the sialylation increases over time in secretor mothers while decreasing in non-secretor. Finally, in combination with the substructure network, we can systematically study glycan synthesis. The product/reactant ratio provides a regular measure indicating the flux through the glycan biosynthetic pathways. Using the HMO dataset, we demonstrate the power of this perspective by showing that more LSTb is converted to DSLNT in the secretor mother using regression modeling. The perspectives made available through GlyCompare are not limited to t-tests and regression models. We have reduced the sparse nature of glycoprofiles, thereby making countless comparisons and new analyses possible.

Conclusions

In conclusion, GlyCompare provides a novel paradigm for describing complex glycoprofiles, thus enabling a wide range of analyses and facilitating the acquisition of detailed insights into the molecular mechanisms controlling all types of glycosylation.

Methods

Data Availability and Implementation

https://github.com/LewisLabUCSD/GlyCompare

Glycoprofile preprocess procedures

There are three procedures required for preprocessing the studied glycoprofiles (**Fig. 1a**). First, glycoprofiles are parsed into glycans with abundance. In each glycoprofile, the glycans are manually drawn and exported with GlycoCT format using the Glytoucan Graphic Input tool[13]. The structure of glycan in GlycoCT format will be further loaded into python using "glypy" (a glycan analysis and glycoinformatics library for Python) for programmatic manipulation. Assuming we have a glycoprofile ***i***, the corresponding abundance of each glycan ***j*** in glycoprofile ***i*** is represented by g_{ij}. For example, the relative m/z peak in the mass spectrum or the abundance value in an HPLC trace, will be rectified by the total abundances of glycans in this glycoprofile $g_{ij}/\Sigma g_{i*}$. Sometime, the glycans with ambiguous topology will be handled by assuming it could belong to every possible structure with the same probability, thereby creating all possible ***n*** structures but with $g_{ij}/n\Sigma g_{i*}$ abundance of each. Second, glycans are annotated with glycan substructure information and this information is transformed into the glycoprofile vector. The substructures within a glycan are exhaustively extracted by breaking down each linkage or a combination of linkages of the studied glycan. Note that this method cannot currently deal with glycans with ring topology. All substructures extracted will be merged into a substructure set ***S***, in which the complexity degrees are sorted and the duplicated substructures are removed. Then, each glycan will match to the substructure set ***S*** resulting in a binary glycan substructure presence (1) or absence (0) vector, named x_{ij}. Lastly, a glycoprofile vector is calculated as $p_i = \Sigma x_{ij} g_{ij}/\Sigma g_{i*}$ representing the abundance of the substructures ***s*** in this glycoprofile, where $p_i = (s_{1i}, \ldots, s_{ni})$. Third, a substructure network is built based on the glycoprofile vectors. The substructure network is a directed acyclic graph and each node denotes a glycan substructure. Given the substructure set S, the root node starts from the monosaccharides or a defined root

core structure, and a child node is a substructure that has only one monosaccharide added to its parent node. We note that one child node might have multiple parent nodes and vice versa. The child node depends on its parent node(s) since it cannot exist alone without any parent node.

Generating the glyco-motif vector bases on the substructure abundance

A larger subset of the substructure network is necessary to uniquely describe a more diverse set of glycoprofiles while fewer substructures are needed to sufficiently describe more similar glycoprofiles. Comparisons become more focused when only examining these variable substructures. By checking the substructure network, the substructures that have exactly the same abundance can be merged without any information loss. In other words, after the substructure network is generated, it is simplified by merging the substructure nodes. The merging criteria are based on how child substructure node s_b depends on the parent substructure node s_a. The dependency is the correlation of their abundance across all glycoprofiles, $corr(s_{a*}, s_{b*})$. If the correlation is 1, we can conclude that the addition of the specific monosaccharide is not perturbed across all glycoprofiles, which means they carry exactly the same information. Thus, the parent node can be pruned without information loss. All remaining nodes, namely, the glyco-motifs, are used to cluster the glycoprofiles.

Meanwhile, the weight of each node is initialized at a value of 1. When a node is removed, the weight is equally divided and distributed to child nodes whose correlation with the removed node is 1. Since this method happens from the root to leaves, the node that serves as checkpoint will gain more weight. The weights $\boldsymbol{W}$ are used later for generating the representative substructures.

Procedures for glycoprofile clustering and identifying representative glycan substructures

The preprocessed glycoprofiles (see details in the "glycoprofile preprocess procedures") generate the glycoprofile vectors to enable further clustering analysis. Here we used the Pearson correlation and 'complete' distance to cluster the glycoprofiles. This procedure clusters the glycoprofiles and substructures.

To identify the representative glycan substructures, a set of glycan substructures with weights $\boldsymbol{W}$ are first aligned. Then, we calculate the sum of monosaccharide weights for each glycan substructure. The representative substructure is thus defined as the glycan substructures with their summed monosaccharide weights greater than 51% of the total weight of glycan substructures. Lastly, the averaged abundances of the representative substructures are generated to assess their differential expressions between different glycoprofiles.

Specification for HMO data

48 milk samples were collected from 6 human mothers (1, 2, 3, 4, 7, 14, 28 and 42 days postpartum (DPP)). HPLC was used to measure the abundances of the 16 major glycans. The relative abundance of each glycan in each milk sample is normalized by $w_i/\Sigma w_*$ for the given sample. Then, we followed the substructure network was built as the procedure mentioned above.

Testing the substructure-phenotype association

The General Equation Estimating (GEE from R::geepack[26,27]) with an exchangeable correlation structure was used to measure associations between motif abundance, secretor status and DPP; these regression models were trained to predict motif abundance from either secretor status or DPP. A Wald Test was used to measure the significance of phenotype contribution.

Product/Reactant Ratio as a proxy for flux and estimating flux-phenotype associations

To further isolate motif-specific effects from biosynthetic biases, we explored methods to control for the reactant-product relations. First, we isolated the relative abundance of motifs separated by only one monosaccharide in the biosynthetic network; these are reactant-product relations like LNT and LSTb. Motif abundance represents the total substructure synthesized therefore when we examine the ratio of product/reactant we measure the total amount of the reactant substructure converted to the product substructure in the sample. The Product/Reactant Ratio is therefore a proxy for flux. Using the same GEE repression modeling approach from testing substructure-phenotype associations, we can measure the association between flux between two glycans and a phenotype of interest.

Acknowledgements

We thanks Joshua Klein for providing generous and timely guidance on using the glypy. This work was conducted with support from the Novo Nordisk Foundation provided to the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517: A.W.T.C.), NIGMS (R35 GM119850: B.B., B.K.P., N.E.L.) and NICHD (R21 HD080682: L.B.).

Author contributions

B.B, B.P.K designed the work. B.B, B.P.K, A.W.T.C., A.K.Y., and N.E.L. performed data analysis. The manuscript was written by B.B, B.P.K, A.W.T.C., L.B., and N.E.L.

Competing interests

The authors declare no competing financial interests.

References

1. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **1**, 90 (2011).
2. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta* **1473**, 4–8 (1999).
3. RodrÍguez, E., Schetters, S. T. T. & van Kooyk, Y. The tumour glyco-code as a novel

immune checkpoint for immunotherapy. *Nat. Rev. Immunol.* **18**, 204–211 (2018).

4. Gutierrez, J. M. *et al.* Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. Preprint at (*bioRxiv*) https://www.biorxiv.org/10.1101/351387 (2018).

5. Gabius, H. J., André, S., Kaltner, H. & Siebert, H. C. The sugar code: functional lectinomics. *Biochimica et Biophysica Acta (BBA)* **1572**, 165–177 (2002).

6. Spahn, P. N. & Lewis, N. E. Systems glycobiology for glycoengineering. *Curr. Opin. Biotechnol.* **30**, 218–224 (2014).

7. Holst, S. *et al.* N-glycosylation profiling of colorectal cancer cell lines reveals association of fucosylation with differentiation and caudal type Homebox 1 (CDX1)/Villin mRNA expression. *Mol. Cell. Proteomics* **15**, 124–140 (2016).

8. Reiding, K. R., Blank, D., Kuijper, D. M., Deelder, A. M. & Wuhrer, M. High-throughput profiling of protein N-glycosylation by MALDI-TOF-MS employing linkage-specific sialic acid esterification. *Anal. Chem.* **86**, 5784–5793 (2014).

9. Yang, Z. *et al.* Engineered CHO cells for production of diverse, homogeneous glycoproteins. *Nat. Biotechnol.* **33**, 842–844 (2015).

10. Anugraham, M. *et al.* Specific glycosylation of membrane proteins in epithelial ovarian cancer cell lines: glycan structures reflect gene expression and DNA methylation status. *Mol. Cell. Proteomics* **13**, 2213–2232 (2014).

11. Čaval, T., Tian, W., Yang, Z., Clausen, H. & Heck, A. J. R. Direct quality control of glycoengineered erythropoietin variants. *Nat. Commun.* **9**, 3342 (2018).

12. Riley, N. M., Hebert, A. S., Westphall, M. S. & Coon, J. J. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun.* **10**, 1311 (2019).

13. Aoki-Kinoshita, K. *et al.* GlyTouCan 1.0--The international glycan structure repository. *Nucleic Acids Res.* **44**, D1237–42 (2016).

14. Campbell, M. P. *et al.* Validation of the curation pipeline of UniCarb-DB: building a global glycan reference MS/MS repository. *Biochim. Biophys. Acta* **1844**, 108–116 (2014).

15. Campbell, M. P. *et al.* UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.* **42**, D215–221 (2014).

16. Rademacher, C. & Paulson, J. C. Glycan fingerprints: calculating diversity in glycan libraries. *ACS Chem. Biol.* **7**, 829–834 (2012).

17. Hosoda, M. *et al.* MCAW-DB: A glycan profile database capturing the ambiguity of glycan recognition patterns. *Carbohydr. Res.* **464**, 44–56 (2018).

18. Alocci, D. *et al.* Understanding the glycome: an interactive view of glycosylation from glycocompositions to glycoepitopes. *Glycobiology* **28**, 349–362 (2018).

19. Klein, J., Carvalho, L. & Zaia, J. Application of network smoothing to glycan LC-MS profiling. *Bioinformatics* **34**, 3511–3518 (2018).

20. Sharapov, S. *et al.* Defining the genetic control of human blood plasma N-glycome using genome-wide association study. Preprint at (*bioRxiv*) https://www.biorxiv.org/10.1101/365486 (2018).

21. Mohammad, M. A., Hadsell, D. L. & Haymond, M. W. Gene regulation of UDP-galactose synthesis and transport: potential rate-limiting processes in initiation of milk production in humans. *Am. J. Physiol. Endocrinol. Metab.* **303**, E365–376 (2012).

22. Hanley, J. A., Negassa, A., Edwardes, M. D. D. & Forrester, J. E. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am. J. Epidemiol.* **157**, 364–375 (2003).

23. Koda, Y., Soejima, M., Liu, Y. & Kimura, H. Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: a fusion gene generated by unequal crossover responsible for the enzyme deficiency. *Am. J. Hum. Genet.* **59**, 343–350 (1996).

24. Kudo, T. *et al.* Molecular genetic analysis of the human Lewis histo-blood group system. II. Secretor gene inactivation by a novel single missense mutation A385T in Japanese nonsecretor individuals. *J. Biol. Chem.* **271**, 9830–9837 (1996).

25. Viverge, D., Grimmonprez, L., Cassanas, G., Bardet, L. & Solere, M. Discriminant carbohydrate components of human milk according to donor secretor types. *J. Pediatr. Gastroenterol. Nutr.* **11**, 365–370 (1990).

26. Yan, J. & Fine, J. Estimating equations for association structures. *Stat. Med.* **23**, 859–874; discussion 875–877, 879–880 (2004).

27. Halekoh, U., Højsgaard, S., Yan, J. The R package geepack for generalized estimating equations. *J. Stat. Softw.* **15**, 1–11 (2006).

Figure Legends

Fig. 1 | The GlyCompare workflow for glycoprofile decomposition and comparison. a and b, A glycoprofile with annotated structure and relative abundance is obtained and the glycans are decomposed to a substructure vector ***S*** and the presence/absence matrix ***X*** is built. Each substructure vector is weighted by the glycan abundance, and the weighted substructure vectors are summed into a glycoprofile vector ***p***. **c**, Seven example glycoprofiles are represented here with their glycoprofile vectors. The glycoprofile vectors are simplified to contain a minimal number of substructures, so a substructure network is constructed to identify the non-redundant glyco-motifs that change in abundance from their precursor substructures. **d**, The glycoprofiles can be re-clustered with simplified glyco-motif vectors for a clearer result. **e**, Clustered substructures can be analyzed to identify the most representative structure in the group. For example, four substructures with different relative abundance were aligned together and the monosaccharides' with weight over 51% was preserved.

Fig. 2 | Glycompare effectively clusters panels of distinct glycoprofiles and identifies differentially expressed representative substructures by resolving sparsity. a, Sixteen glycoprofiles from glycoengineered recombinant EPO cluster poorly when based solely on raw glycan abundance. **b**, GlyCompare was used to compute and cluster EPO glyco-motif vectors. This resulted in three dominant clusters of glycoprofiles and a few individuals that have very different structural pattern (distance threshold=0.5) and twenty-four clusters of glycan substructure (distance threshold=0.19). **c**, The representative substructure table contains representative substructures for each of the 24 substructure clusters. The color scale represents the averaged abundances of the substructures in each cluster. The substructures are sorted based

on the glycan structure complexity, followed by the number of branches, the degree of galactosylation, sialylation, and fucosylation. **d**, The significantly differentially expressed glycan substructures are illustrated by z-score normalized abundance of twenty-four glycan substructures, compared with WT.

Fig. 3 | Changes in representative substructures are quantified and compared to WT with the standardized abundance bar plot. a, Differential fucosylation is illustrated for the Fut8 knockout. The red (black) triangles represent the presence/absence of fucose in the representative substructures. Differential sialylation is illustrated for the St3gal4/6 knockout. The purple/black diamonds represent the presence/absence of the sialylation in the representative substructures. **b**, Changes in branching are presented for the Mgat4A/4B/5 knockouts. The tetra-antennary substructures (Rep16-Rep22) decreased considerably. The triantennary substructures with elongated GlcNac (Rep13-Rep14) increase significantly (p-value < 0.01). However, the elongated triantennary structure (Rep15) decreases significantly for the Mgat5 and Mgat4B knockouts (p-value < 0.001). In the CHO dataset, the glycan substructure generated by Mgat4A/4B and Mgat5 will be considered as the same topologically.

Fig. 4 | Analysis of intermediate substructures with GlyCompare elucidates associations with secretor status over time, which are missed in the standard whole-glycan analysis. a, The substructure intermediates for four connected HMOs are shown here. The synthesis of larger HMOs must pass through intermediate substructures that are also observed HMOs, where the substructures are as associated with measured HMOs as follow X38=LNT, X80=LSTb, X103=DSLNT, X138=DSLNH. **b-e**, Over time (DPP), X80, LSTb, DSLNT, and DSLNH show

different trends for secretors and non-secretors. Furthermore, the abundance of aggregated X80 shows significant positive-correlation with secretor and negative-correlation with non-secretor.

Fig. 5 | Product-Reactant Ratio of two reactions. X38, the LNT substructure, is a precursor to X80, the LSTb substructure, which is a precursor to X103, the DSLNT substructure. We estimate the flux of these conversions from X38 to X80 and X80 to X103 by examining the product/reactant ratio; the proportion of the total synthesize reactant converted to the product. **a**, LSTb/LNT motif relative abundance ratios are not associated with secretor status while **b**, DSLNT/LSTb ratios are. Each line is a simple linear regression leveraging sample-specific data.

Fig. 1 | The GlyCompare workflow for glycoprofile decomposition and comparison.

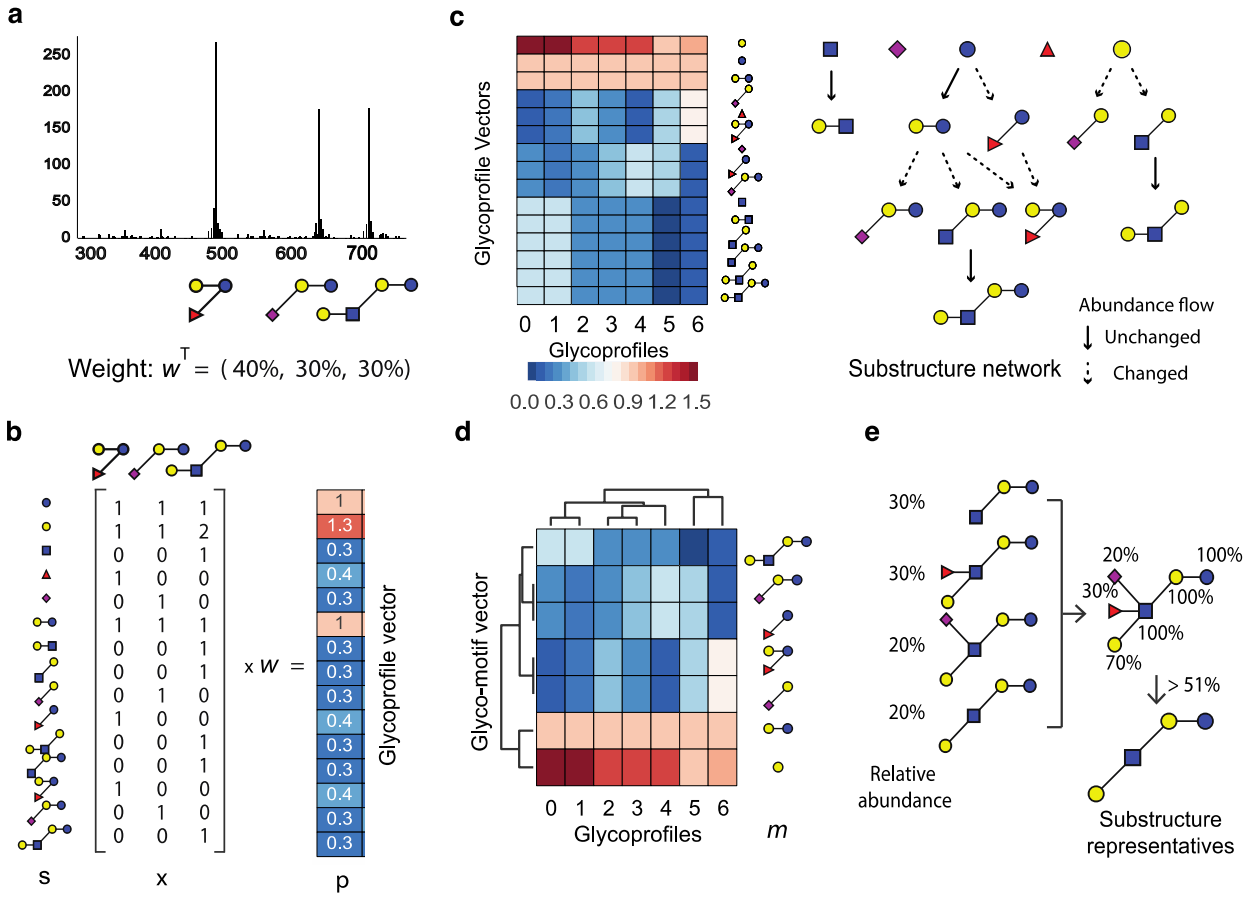

Fig. 2 | Glycompare effectively clusters panels of distinct glycoprofiles and identifies differentially expressed representative substructures by resolving sparsity.

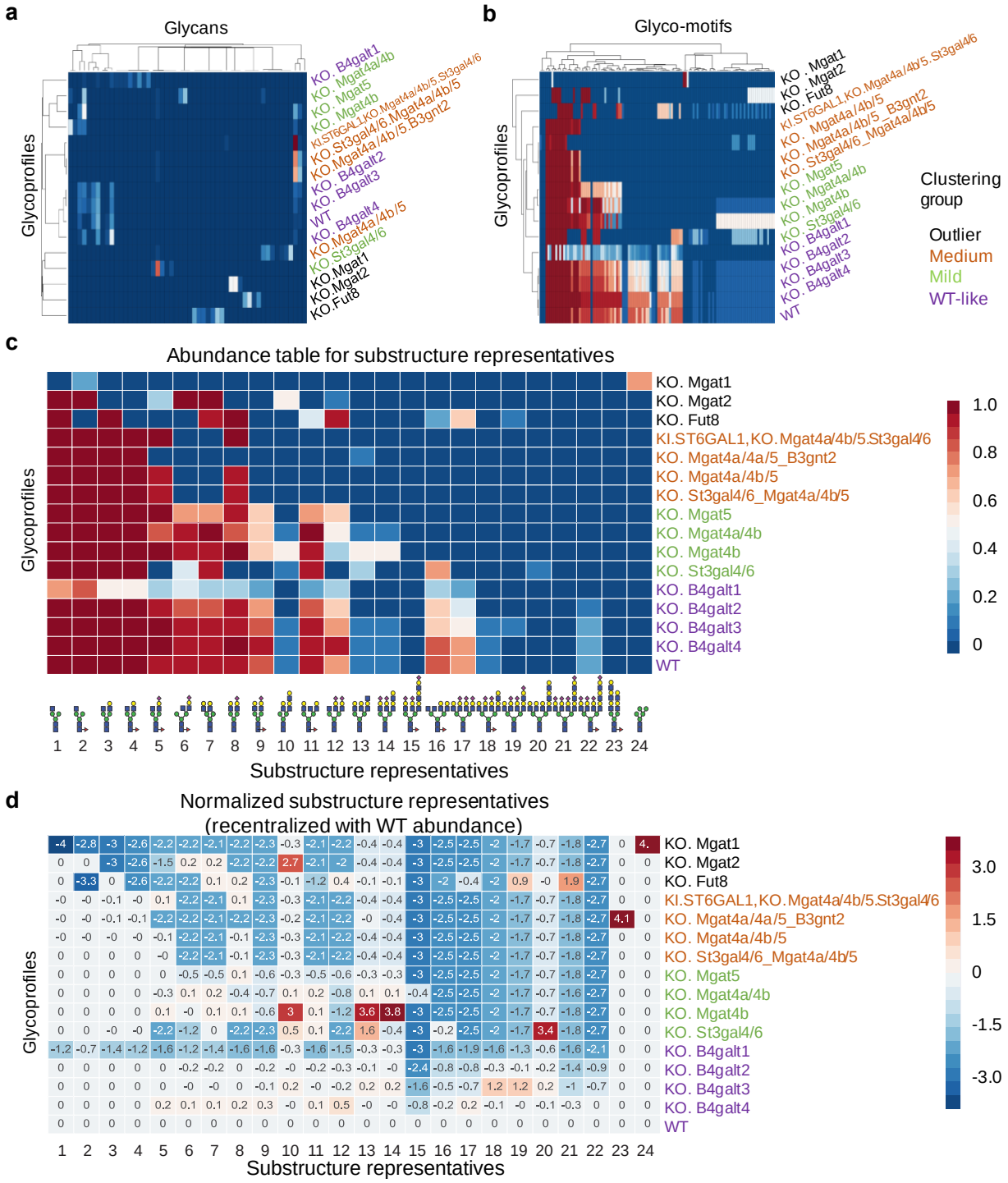

Fig. 3 | Changes in representative substructures can be quantified and compared to WT with the standardized abundance bar plot.

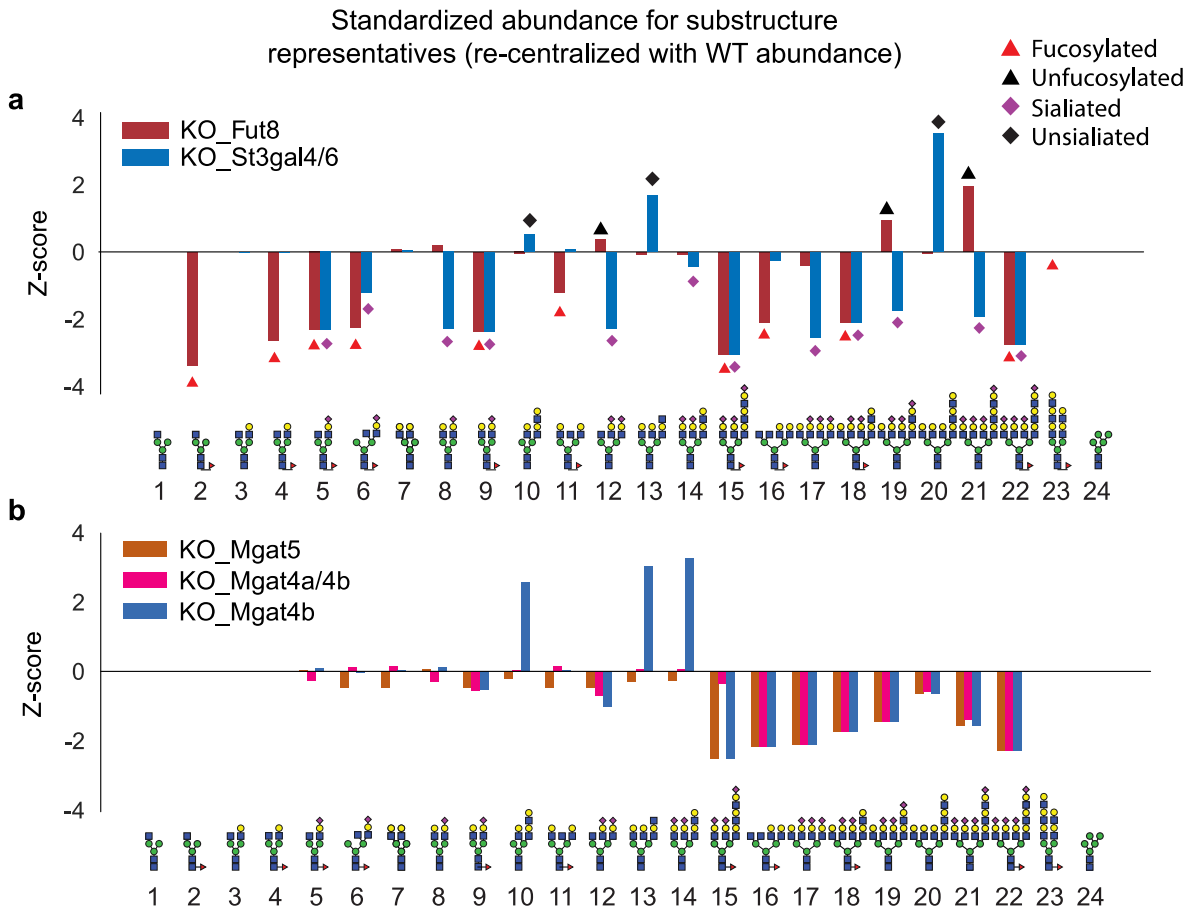

Fig. 4 | Analysis of intermediate substructures with GlyCompare elucidates associations with secretor status over time, which are missed in the standard whole-glycan analysis.

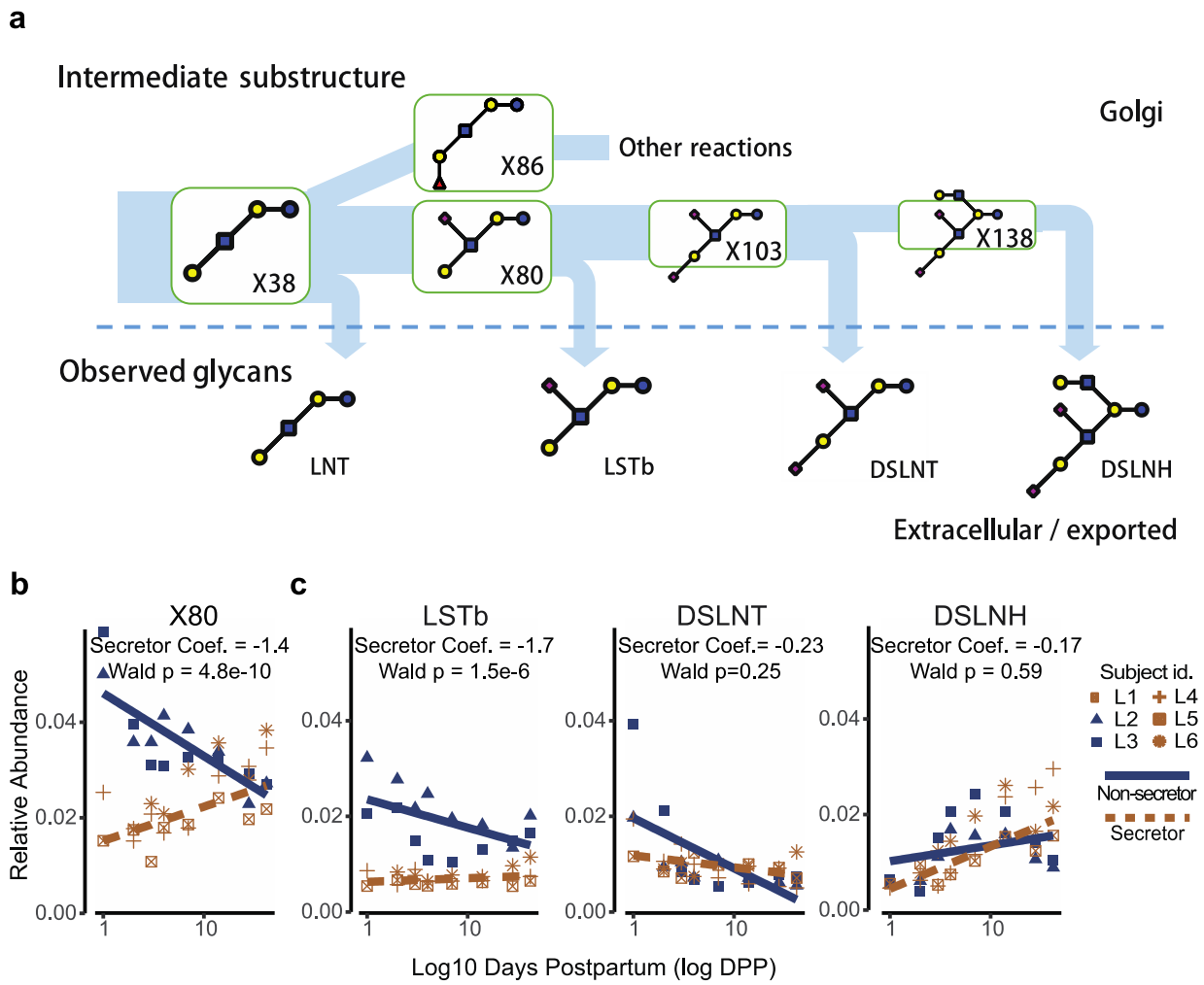
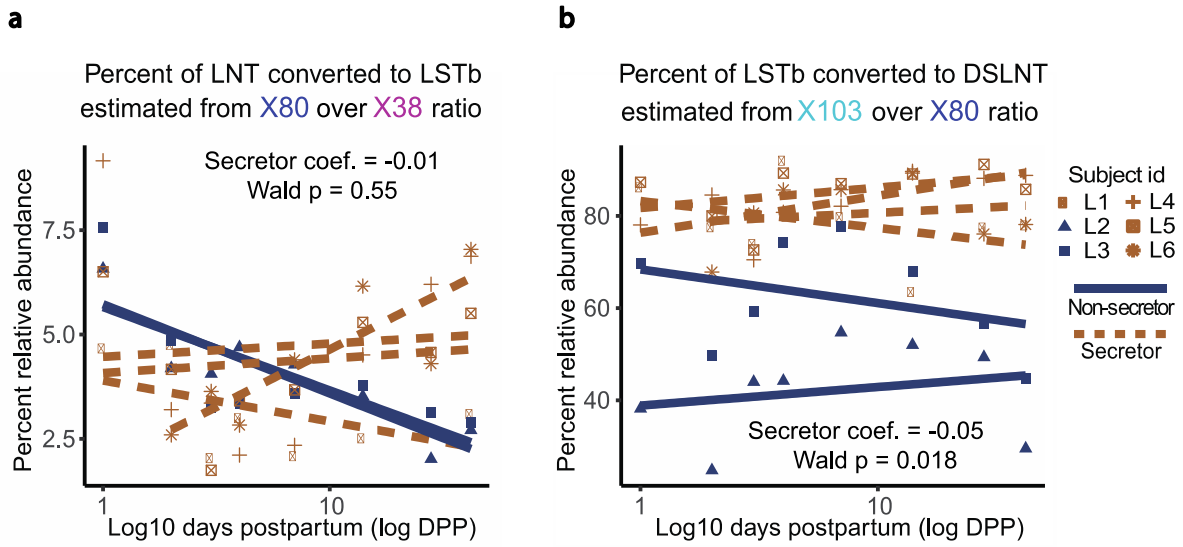
Fig. 5 | Product-Reactant Ratio of two reactions.