NAR Genomics and Bioinformatics, 2024, **6**, lqae169
https://doi.org/10.1093/nargab/lqae169
Advance access publication date: 18 December 2024
Standard Article

Cell- and tissue-specific glycosylation pathways informed by single-cell transcriptomics

Panagiotis Chrysinas[1,†], Shriramprasad Venkatesan[1,†], Isaac Ang[2], Vishnu Ghosh[1], Changyou Chen[3], Sriram Neelamegham[1,*] and Rudiyanto Gunawan[1,*]

[1]Department of Chemical and Biological Engineering, University at Buffalo-SUNY, 308 Furnas Hall, Buffalo, NY 14260, USA
[2]Department of Computer Science, University of Illinois Urbana-Champaign, 201 North Goodwin Avenue, Urbana, IL 61801, USA
[3]Department of Computer Science and Engineering, University at Buffalo-SUNY, 338 Davis Hall, Buffalo, NY 14260, USA

[*]To whom correspondence should be addressed. Tel: +1 716 645 0952; Fax: +1 716 645 3822; Email: rgunawan@buffalo.edu
Correspondence may also be addressed to Sriram Neelamegham. Email: neel@buffalo.edu
[†]The first two authors should be regarded as Joint First Authors.

Abstract

While single-cell studies have made significant impacts in various subfields of biology, they lag in the Glycosciences. To address this gap, we analyzed single-cell glycogene expressions in the Tabula Sapiens dataset of human tissues and cell types using a recent glycosylation-specific gene ontology (GlycoEnzOnto). At the median sequencing (count) depth, ∼40–50 out of 400 glycogenes were detected in individual cells. Upon increasing the sequencing depth, the number of detectable glycogenes saturates at ∼200 glycogenes, suggesting that the average human cell expresses about half of the glycogene repertoire. Hierarchies in glycogene and glycopathway expressions emerged from our analysis: nucleotide-sugar synthesis and transport exhibited the highest gene expressions, followed by genes for core enzymes, glycan modification and extensions, and finally terminal modifications. Interestingly, the same cell types showed variable glycopathway expressions based on their organ or tissue origin, suggesting nuanced cell- and tissue-specific glycosylation patterns. Probing deeper into the transcription factors (TFs) of glycogenes, we identified distinct groupings of TFs controlling different aspects of glycosylation: core biosynthesis, terminal modifications, etc. We present webtools to explore the interconnections across glycogenes, glycopathways and TFs regulating glycosylation in human cell/tissue types. Overall, the study presents an overview of glycosylation across multiple human organ systems.

Introduction

Glycosylation is a ubiquitous post-translational modification that results in the formation of an array of cellular complex carbohydrate structures or glycans (1). These glycans, which appear either in branched or extended form on the cell surface or as single-monosaccharide additions within cells, control or fine-tune a multitude of biological functions during normal physiology and disease (2,3). The common glycoconjugate types on mammalian cells include the branched N-linked glycans on glycoproteins, O-GalNAc (N-acetyl galactosamine) type O-glycan modifications on glycoproteins, long repeating saccharide chains called glycosaminoglycans (GAGs) on a select set of proteoglycans, carbohydrate modifications on glycolipids, and finally O-GlcNAc type single residue modifications on nuclear proteins and transcription factors (TFs). Besides these major families of glycoconjugates, glycans also form the anchor for glycosylphosphatidylinositol (GPI)-linked cell-surface proteins. There also exist a growing list of rarer O-linked glycan modifications, including O-Glc (glucose), O-Fuc (fucose) and O-Man (mannose) type glycosylation (4).

Glycans on cells are formed by the concerted action of ∼2% of the expressed proteome that are collectively called 'glycoEnzymes.' These enzymes are products of the corresponding 'glycogenes.' An ontology called 'GlycoEnzOnto' has recently been curated to organize the existing knowledge of human glycoEnzymes within the domain of Glycosciences (5). In this ontology, the ∼400 glycogenes are annotated according to their molecular functions, biological processes and physical location (cellular component), following the Gene Ontology convention (6). Besides the genes encoding enzymes in the glycan biosynthesis, GlycoEnzOnto also includes the entities involved in the regulation of nucleotide-sugar metabolism, glycosyl-substrate/donor transport, glycan degradation and other regulatory components. From the molecular function perspective, glycogenes are grouped into glycosyltransferases, other transferases (e.g. sulfotransferases), modifying enzymes (e.g. epimerases and kinases), glycosidases, molecular transporters and other regulators. Additionally, glycogenes/glycoEnzymes can also be classified according to their role in glycoconjugate biosynthesis, including (i) the 'initiation' step that results in the attachment of the first monosaccharide or oligosaccharide to the protein/lipid, (ii) the 'elongation and branching' reactions that extend the original glycan often via lactosamine chain synthesis/branching, and (iii) the 'termination or capping' processes that prevent further chain extension. GlycoEnzOnto and other databases such as GlycoGene Database (GGDB) in GlyCosmos portal

Received: July 18, 2024. **Revised:** November 6, 2024. **Editorial Decision:** November 11, 2024. **Accepted:** November 21, 2024

(7) represent invaluable shared resources for Systems Glycobiology analysis (8–12).

Advances in single-cell profiling technologies have generated voluminous multimodal single-cell data and have transformed our understanding of cell biology, from development (13,14) to immune systems (15,16) and to aging (17,18). Of note is the scRNA-seq dataset for human from the Tabula Sapiens (TS) project, comprising 483 152 human cells from 15 donors that are organized into 24 tissues and over 400 cell types (19). Such single-cell data may reveal subtle, yet potentially vital, differences in glycosylation processes among different cells within the same tissue or organ as this is not possible using bulk sequencing. A study by Joshi *et al*. employed the TS dataset to examine the activity and regulation of glycosyltransferases and associated pathways across human tissues and cell types. The study identified key transcriptional regulatory hotspots in different glycopathways and generated a tool called Glycopacity for characterizing the capacity of glycosyltransferase pathways using gene expression data (20). Besides gene expression, other recent studies generated multimodal single-cell data of lectin-based profiling and scRNA-seq for integrative analysis of glycans (21–23). While different single-cell methods have their advantages and disadvantages, they share a few issues, such as low messenger RNA (mRNA) capture efficiency and high dropout rates, that particularly affect the measurement of genes with low expression (24,25).

This study presents a comprehensive analysis of single-cell glycogene and glycopathway expressions in the TS dataset, informed by GlycoEnzOnto, to shed light on the variation of glycopathway expression across various cell and tissue types. Our results show that only ∼50% of glycogenes are expressed in a given cell, with expression levels varying depending on gene function. In contrast to conventional thinking based on microarray/quantitative-PCR data analysis (26) that suggests that glycogenes are lowly expressed, our more holistic single-cell analysis reveals that the glycogenes are expressed at levels that are comparable to other protein-coding (PC) genes. Further, our analysis presents a map of glycopathway expression across tissue, illustrating the inherent heterogeneity across human cells and tissues. Specifically, the findings showed how enzymes involved in the metabolism of nucleotide sugars (NSs), glycan degradation processes and biosynthesis of core structures exhibit uniform and ubiquitous expression patterns across cell types and tissues, consistent with their foundational roles in glycan biosynthesis. Meanwhile, terminal glycoenzymes often serve specialized roles, and they are more selectively expressed in individual cell types. Lastly, the analysis of transcriptional factors using mutual information (MI) of TF–glycogene expression in the TS dataset fills the gap in knowledge of transcriptional regulators of glycosylation (27). The result reveals five regulatory modules (RMs), with each module controlling a different aspect of glycosylation. To bolster accessibility, we also developed webtools, called glycoCARTA and glycoTF (links available at virtualglycome.org), to allow further exploration of glycogenes, glycopathways and related TFs at single cell level.

Materials and methods

Data preprocessing

The scRNA-seq data in the TS project were generated using two different single-cell sequencing technologies: 10X and Smart-seq, with the majority of the data coming from 10X (456 101 cells versus 27 051 cells). TS scRNA-seq data were obtained from the public website (28). This study focused only on scRNA-seq data from 10X platform to avoid any potential batch effects associated with different sequencing platforms.

The data preprocessing is illustrated in Figure 1A, with individual steps being carried out using the Python package scanpy (29). The analysis started with the decontaminated Unique Molecular Identifier (UMI) counts from the TS dataset. Decontamination of background RNA was previously performed using the method decontX (30). Following the standard practice, UMI counts were scaled cellwise so that each cell has a count depth of 10 000. This scaling produced relative RNA abundances x_i that are comparable across cells.

For differential expression (DE) analysis, the scaled UMI counts were log-transformed (i.e., log $(x+1)$) to satisfy the input requirement of the method MAST (31). For subsequent analyses, we subset the preprocessed count matrix to 19 847 transcripts (from a total of 58 559) associated with PC genes as defined in BioMart (access date 10 October 2022). On average, PC genes make up 89.4% of the total number of reads. Among the PC genes, we extracted data for 398 glycosylation-related genes (glycogenes, see Supplementary Table S2) as defined in the GlycoEnzOnto (5). Python and R codes and the list of PC genes used for data analysis are available from 10.5281/zenodo.14177056.

scVI + UMAP embedding

To visualize glycogene expression in single cells, single-cell variational inference (scVI) was applied to the decontaminated glycogene UMI counts to generate a lower dimensional embedding of the glycogene expression (32). The method scVI produces a probabilistic latent space of single-cell gene expression data based on zero-inflated negative binomial distribution. Specifically, the Python's scvi-tools package (33) was implemented with the following parameters: $n_{\text{latent}} = 50$, $n_{\text{layers}} = 3$, $n_{\text{latent}} = 50$ and dropout rate = 50. For this study, only the variational posterior of the scVI model was employed. For visualization of this embedding, a Uniform Manifold Approximation and Projection (UMAP) using $n_{\text{neighbors}} = 15$ and $n_{\text{components}} = 2$ (2D) was used (34).

Differential expression analysis of glycopathways

Glycosylation-related pathways (glycopathways) are described in Supplementary Table S3. DE analysis using MAST requires as inputs the $\log(x+1)$-transformed UMI counts for the glycogene expression. For each pathway, its expression was evaluated by averaging the expression of glycogenes belonging to that pathway. DE analysis of glycopathways was then performed using the MAST (model-based analysis of single-cell transcriptomics). MAST adapts a hurdle model to tackle zero-inflation and biomodality of single-cell transcriptome data (31). As illustrated in Figure 2A, MAST produces fold-change differences of the mean expression of a glycopathway between cells from a specific tissue, with respect to all other tissue with the same glycopathway. Associated statistical significance is also calculated using *P*-values adjusted for multiple hypothesis testing using Bonferroni correction (35). The DE analysis was implemented using the Seurat package (version 4.1.0) in R, specifically *FindAllMarkers* function (36). The code and the result of DE analysis using MAST is available from 10.5281/zenodo.14177056.

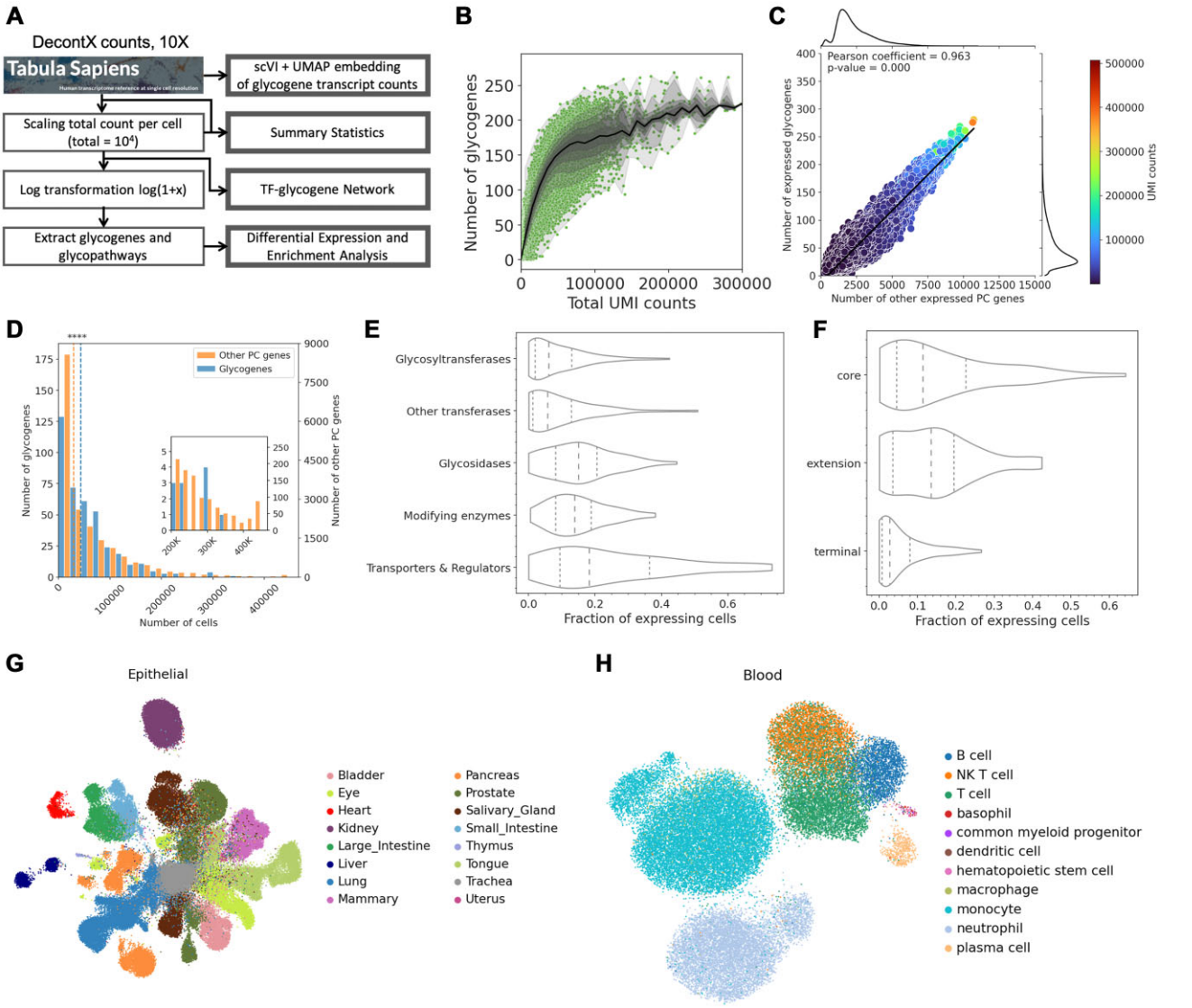

Figure 1. Transcriptomic analysis of glycogenes at single cell level. (**A**) Summary of data preprocessing and analyses of TS data (see 'Materials and methods' section for details). (**B**) Number of glycogenes with RNA count > 0. Each dot represents a cell. Dark line represents the median value. The shaded areas show the contours of percentiles at 10% increments from the median, spanning the 0th to 100th percentile. The data collection results in a shoulder at ∼65 000 UMI counts/cell corresponding to ∼165 glycogenes. (**C**) A positive correlation is observed between the number of detected glycogenes versus other PC genes. The face color represents the UMI depth. (**D**) Distribution of PC and glycogene expressions in terms of the number of expressing cells (i.e., cells with nonzero RNA count for the gene). Glycogenes are generally more commonly expressed in the TS cells than other PC genes. (**E**) Distribution of single cell expression of glycogenes. Glycogenes are grouped based on their biological functions as defined in the GlycoEnzOnto (see Supplementary Table S2). 'Transporters and regulators' are generally more broadly expressed compared to other glycogenes including glycosyltransferases. (**F**) Distribution of single cell expression of glycosyltransferases in core, extension and terminal groups (see Supplementary Table S3). Core enzyme expression is higher compared to extension and terminal modifiers. (**G–H**) UMAP visualization of scVI latent embedding of glycogene expression in epithelial cells and blood tissue.

Glycopathway enrichment analysis

Enrichment analysis was performed to assess whether cells from a specific tissue are over-represented or depleted with cells expressing a given glycopathway. Here, a cell is labeled as an 'expressing cell' when the average expression (scaled UMI count) of genes in a glycopathway is nonzero. For a given pair of glycopathway and tissue, a contingency table as shown in Table 1 was constructed to distribute the cells into two distinct categorizations: expressing cells/non-expressing cells and cells in the tissue/cells not in the tissue. The odds ratio OR, given by ad/bc, indicates the over-representation (odds ratio > 1 or log(OR) > 0) or depletion (odds ratio < 1 or log(OR) < 0) of expressing cells in a tissue. The statistical significance was established via Fisher's exact test based on hypergeometric sampling. The enrichment analysis was implemented in Python using *fisher_exact* function from the scipy package (version 1.10.1).

Table 1. Contingency table for glycopathway enrichment analysis

	Expressing cells	Non-expressing cells
Cells in tissue	*a*	*b*
Cells not in tissue	*c*	*d*

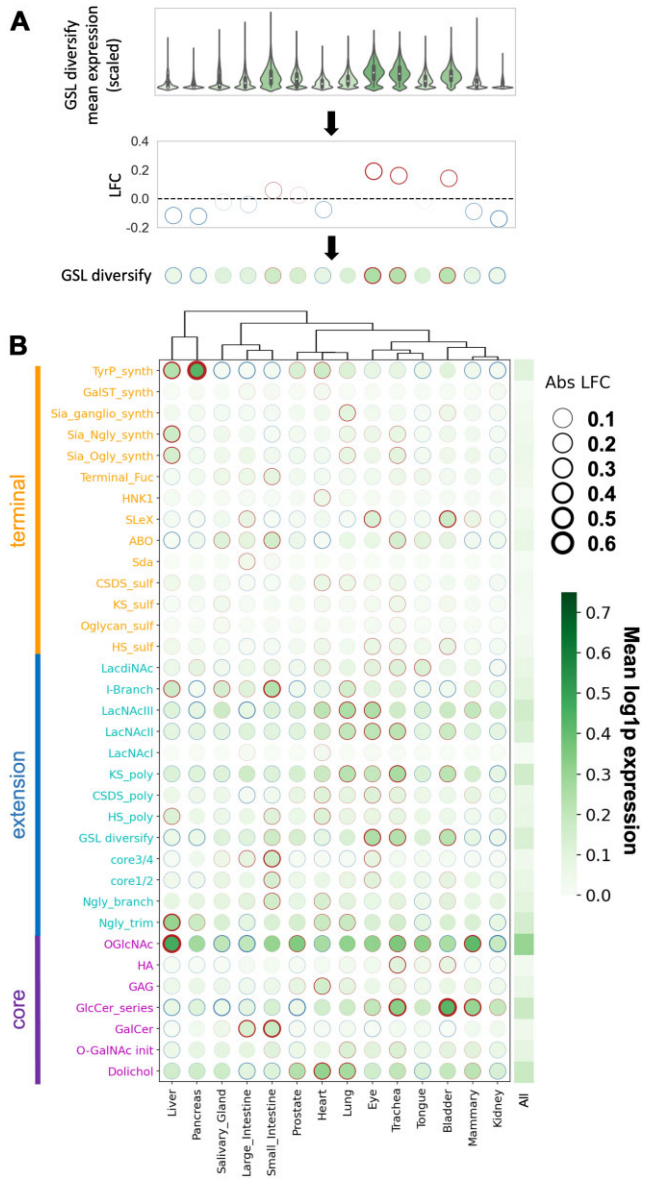

Figure 2. Differential expression analysis of glycopathway in epithelial cells. (**A**) 'GSLs diversify' pathway including the gene set B4GALNT1, B3GALT4, B3GNT5, B3GALT5, B4GALT1, B3GALNT1 and A4GALT is used to illustrate the calculation scheme. Here, mean pathway gene expression in each tissue is first calculated from the zero-inflated single-cell data. LogFC expression is then determined, and this is presented using dots where a thicker red (blue) linewidth represents higher (lower) levels of the pathway expression in a given tissue with respect to all other tissue. The intensity of the face color of the dot represents the mean expression of the glycopathway among the cells in a specific tissue. (**B**) DE of glycopathways is presented for selected core, extension and terminal pathways for epithelial cells. The grayscale heatmap in the last column presents the mean glycopathway expression among all epithelial cells. As an example, O-GlcNAc-related genes (OGA and OGT) are highly expressed across tissue. Among the tissue, this is most highly expressed in the liver compared to kidney and large intestine. In contrast the GSL diversity genes are higher in eye and trachea compared to other tissue. Mean expressions data are provided in Supplementary Table S4, while logFC data are available from online repository 10.5281/zenodo.14177056.

Transcriptional factor analysis

The curation of TF–glycogene and glycopathway interactions involved evaluating MI of single-cell expression between every possible pair of TF–glycogene in the TFLink database (37). TF–gene interactions in the TFLink database were originally compiled from numerous databases that relied on different evidence of TF binding on the regulatory elements of the genes (37). Here, MI was used to provide additional evidence for TF–glycogene interactions based on shared information in their single-cell expression. The following equation gives the basis for evaluating MI for TF–glycogene relations:

$$I(\mathrm{TF},\ \mathrm{G}) = \sum_{t_i}\sum_{g_j} P_{\mathrm{TF},\ \mathrm{G}}(t f, g)\log\frac{P_{\mathrm{TF},\ \mathrm{G}}(t f, g)}{P_{\mathrm{TF}}(t f)\,P_{\mathrm{G}}(g)} \quad (1)$$

where $P_{\mathrm{TF},\ \mathrm{G}}$ denotes the joint probability distribution of single-cell expression of a transcription factor TF and a glycogene G, and P_{TF} and P_{G} denote the marginal probability distribution of TF and G, respectively. We evaluated $I(\mathrm{TF},\ \mathrm{G})$ using the scaled UMI counts from all 10X cells in the TS dataset. The calculation of MI was performed using the Python package sklearn (38).

We employed the TFLink database (37) as the ground truth to assess the accuracy of MI scores for establishing TF–glycogene interactions. Specifically, we evaluated the area under precision-recall curve (AUPRC) using the Python package sklearn (38). The AUPRC has a value between 0 and 1 with 1 describing the ideal predictor. To assess the empirical P-value, we performed a bootstrap approach by generating a set of random MI scores ($n = 100\ 000$), representing the outputs of a random predictor. The empirical P-value is set to the proportion of the AUPRCs from a random predictor that is higher than the AUPRC of the MI scores computed using the single-cell gene expression data.

Transcriptional regulatory module analysis

Transcriptional RMs were identified by hierarchical clustering of glycogenes using their log1p of MI with TFs, i.e., $\log(1 + I(\mathrm{TF},\ \mathrm{G}))$ (see Transcription factor analysis). Thus, each cluster corresponds to a group of glycogenes with similar MI patterns. The hierarchical clustering was performed using the 'linkage' function from the SciPy library. Specifically, we employed the complete linkage method which defines the distance between two clusters as the maximum distance between any pair of elements across the clusters. The Euclidean distance metric was used to quantify dissimilarity between glycogenes. The resulting linkage matrix was visualized as a dendrogram. Finally, the clusters of glycogenes were assigned using the *fcluster* function from the SciPy library with the criterion 'maxclust' to obtain the desired number of clusters ($n = 5$). The hierarchical clustering above was implemented using the scikit-learn package (version 1.2.2).

Each glycogene cluster was taken as a Transcriptional Regulatory Module (TRM). Enrichment analysis was performed using the Fisher's exact test to identify the over-representation of glycopathways. For this purpose, we employed the glycopathway definition given in Supplementary Table S3. Given two sets of glycogenes, one from a TRM and another from a glycopathway, we constructed the contingency table as shown in Table 2. The odds ratio calculation ($OR = mt/ns$) and Fisher's exact test were performed following the same procedure for the ED analysis (see 'Glycopathway enrichment analysis' section). Hierarchical and k-means clustering were im-

Table 2. Contingency table for TRM enrichment analysis

	Glycogenes in pathway	Glycogenes not in pathway
Glycogenes in TRM	*m*	*n*
Glycogenes not in TRM	*s*	*t*

plemented using the scikit-learn package (version 1.2.2) and the Fisher's exact test using the scipy package (version 1.10.1) in Python. The statistical significance was established using Benjamini–Hochberg adjusted *P*-values to account for multiple hypothesis tests (39).

To identify the key TFs, we developed a ranking procedure that generates a list of TFs, ordered based on their relevance to each TRM. First, for every glycogene within a cluster, we sorted the TFs according to their MI score with the glycogene. Then, we calculated an average ranking for each TF across all glycogenes within the cluster. This process was replicated for each glycogene cluster. The results of this analysis are presented in Supplementary Table S6.

Results

Glycogene expression in single cells

As glycogenes are traditionally thought to be lowly expressed (26), this study assessed the ability of single-cell data to inform us about genes and pathways involved in glycosylation by analyzing scRNA-seq data from the TS project. We applied the bioinformatics analysis workflow depicted in Figure 1A to the TS dataset, focusing on glycogenes and glycopathways presented in GlycoEnzOnto (5). We compared the expression profiles of 400 glycogenes to 19 447 PC genes in the TS dataset (Supplementary Table S1). As expected, the number of detectable glycogenes in the TS cells increased with the cell's UMI count (Figure 1B). Specifically, at the median (mean) UMI count depth of 6496 (10 181), we detected between 2 and 98 glycogenes (2 to 118 genes) with RNA count > 0 and a median value of 40 glycogenes/cell (mean value of 58 glycogenes/cell). Thus, only 10% of all glycogenes are detected in the median cell in the TS dataset (14.5% of glycogenes for the mean cell). The low level of detection of glycogenes (i.e., 10–14.5%) might stem from the overall low expression of glycogenes. In this regard, increasing UMI count depth did enhance glycogene detectability. However, the maximum number of captured glycogenes reached a plateau at ∼220, suggesting that, at most, only 50–60% of all glycogenes are expressed in individual cells. Lastly, only a marginal improvement in the number of detected glycogenes was observed beyond ∼65 000 UMI counts per cell, at which point about 165 glycogenes (median) were detected (Figure 1B).

We observed a strong correlation between the number of glycogenes and the number for other PC genes detected in cells (Figure 1C). Comparing the distribution of expression between glycogenes and other PC genes, using the fraction of expressing cells as an indicator of gene expression level, revealed a significant difference between the two groups (*P*-value $= 1.19 \times 10^{-7}$, Kolmogorov–Smirnov test) (see 'Materials and methods' section and Figure 1D). Interestingly, in the TS cells, glycogenes were more commonly expressed than other PC genes (*P*-value $= 5.33 \times 10^{-5}$, two-sided Wilcoxon rank sum test). This trend is consistent across different tissues (see Supplementary Figure S1). But, despite this prevalence, glycogenes were not among the highly expressed genes in the TS dataset (Figure 1D inset)—in fact, glycogenes were depleted among the top 10% of highest expressing PC genes (odds ratio $= 0.655$, *P*-value $= 0.03$, two-sided Fisher's exact test). Overall, the data suggested that while the glycogenes may not be among the most highly expressed genes, they are ubiquitously expressed commonly at levels comparable to or higher than the average PC gene.

Variability in glycogene expression patterns across cell types and tissues

We delved deeper into the variability of glycogene expression among functional sub-groups using the GlycoEnzOnto as a guide. To do this, we evaluated the fraction of cells expressing glycogenes across different sub-groups (Supplementary Tables S2 and S3). Figure 1E reveals that glycogenes belonging to the 'Transporters and Regulators' sub-groups—that is, genes involved in the creation of nucleotide-sugars, monosaccharide transport and related metabolism—generally exhibited higher expression than other glycogene sub-groups. Glycogenes responsible for glycan modifications and those producing glycosidases displayed comparable expression levels with both gene groups presenting moderate expression. Finally, the glycotransferases and other transferases demonstrated the lowest expression levels. Delving further into glycogenes associated with the biosynthesis of core structures (core), glycan chain elongation (extension) and capping of glycan structures (terminal), the core and extension groups showed similar levels of expression that were higher than the expression of glycogenes in the terminal group (Figure 1F). These patterns align with the role of the core enzymes in initiating the formation of specific glycan types, except perhaps for the case of O-GalNAc type carbohydrate chain formation that can be catalyzed by various isoenzymes. Thus, these core genes are more broadly expressed in various cells compared to terminal modifiers that are expressed in a tissue specific manner (20). Although our analysis employed the fraction of expressing cells as the metric for gene expression level—following the recommendation for lowly expressed genes (40)—we observed similar trends using the mean expression of genes across cells (see Supplementary Figure S2).

Next, we investigated how glycogene expression pattern in individual cells varies across different cell types and tissue types using scVI for latent embedding and UMAP (Unified Manifold Approximation and Projection) for 2D visualization (Figure 1G and H; Supplementary Figure S3). Examining epithelial cells (Figure 1G), clusters (grouping) of cells emerged in the UMAP plot following their tissue sources. Interestingly, even for cells from the same tissue, for example liver, pancreas and salivary gland, distinct cell groupings appeared. We made similar observations for endothelial, stromal and immune cells in the TS dataset (see Supplementary Figure S3A and B). Shifting focus to cells in the blood tissue (Figure 1H and Supplementary Figure S3D), these cells formed clusters according to their lineage along the hematopoietic stem cell (HSC) differentiation pathway. Specifically, cells from the lymphoid path, including B cells, T cells, Natural Killer (NK) T cells and plasma cells appeared in overlapping clusters, while cells from the granulocyte—macrophage lineage (macrophages, monocytes and neutrophils) formed separate groups. The overt grouping of cells, influenced by their tissue of origin and lineage, suggests that mammalian tissues and

cell types possess unique single-cell glycogene expression patterns, potentially indicating their varied glycan structures. In subsequent analyses, as 2D UMAP plots may distort cell-cell similarities in single-cell gene expression (41,42), we verified our observations on glycogene expression directly, without relying on UMAP latent embeddings.

Glycopathway expressions vary with tissue and cell type of origin

A rich diversity of glycan structures arise from sets of reactions operating together as 'glycopathways'. To gauge the expression of these glycopathways in TS cells, we calculated the expression of glycopathways in the TS cells by taking the average expression of the genes from each glycopathway as delineated in the GlycoEnzOnto (see Supplementary Table S3). To further discern the patterns of glycopathway expressions, DE analysis was performed (Figure 2). As illustrated in Figure 2A, the DE analysis combined two information: zero-inflated mean expression of glycopathways for cells in each tissue type and log2-fold change (logFC) of the glycopathway mean expression value in a given tissue against cells in all other tissues. The intensity of green color in the DE heatmap plot informs the glycopathway expression while the border thickness indicates the logFC. A higher (lower) glycopathway expression suggests a higher (lower) capacity of the related glycan processing (20). In the example of enzymes involved in extension of the glycosphingolipid (GSL) core and its diversification into ganglio-, lacto-, neolacto- and globo-series (GSL diversify, Figure 2A), we found a higher expression of relevant genes in eyes, trachea and bladder compared to other tissues.

Figure 2B visualizes the mean and DE of various glycopathways across epithelial cells found in multiple tissue-types. Here, glycopathways are grouped into three major functional categories: core, extension and terminal pathways (5). The results for three additional glycopathway groups: core subclass, NS metabolism and degradation processes are provided in Supplementary Figure S4A. The same analyses are also performed for endothelial, stromal and immune cell types, and the results are presented in Supplementary Figure S5A and S7A. Supplementary Table S4 provides the average expressions of the glycopathways for the cell types in the TS dataset. The color scale bar for mean expression used in Figure 2 and the aforementioned Supplementary figures is the same, allowing direct comparison between the cell/tissue types. Additionally, the observations are independent of cell sample size since the mean expression of glycogenes in a system did not depend on either the number of cells in the tissues ($\rho = -0.22$, P-value $= 0.30$) or the number of cells of a given population ($\rho = -0.15$, P-value $= 0.81$).

The data present several striking observations. Comparing the average expressions across cell types, endothelial and stromal cells consistently manifested higher levels of glycopathway expression in comparison to epithelial and immune cells (see Supplementary Table S4). Surveying the groups of glycopathways, the core, NS metabolism and degradation groups had the highest mean glycogene expressions. These observations are in agreement with the expression analysis of glycogenes from these pathways in Figure 1E and F. The trend also underscores the broad functions that these groups of glycopathways have in terms of controlling global glycan turnover rates and pathway initiation steps. In general, the most highly expressed core gene set across all cell/tissue type belonged to the O-GlcNAc forming enzymes OGT and OGA. This highlights the importance of O-GlcNAc post-translational modification in regulating a broad swath of cellular signaling, transcription and disease processes (43). Besides these enzymes, we also observed consistent high expressions of several other core-pathways in diverse cell types, particularly those initiating GSL biosynthesis (i.e., 'GlcCer-series') and those initiating the synthesis of N-linked glycans (i.e., 'dolichol pathways'). In addition to the core dolichol pathway, high gene expression was also observed for N-glycosylation processing enzymes that trim the initial dolichol precursor to enable protein folding and the biosynthesis of complex type structures. Finally, the prevalence of core-3 and core-4 O-GalNAc biosynthetic genes were largely restricted to the intestines, and this too is consistent with literature knowledge (44,45).

GAGs were expressed in stromal cells, particularly with respect to hyaluronan forming enzymes (HAS1, HAS2 and HAS3), which were consistently high in fibroblasts and connective tissue in multiple organs. Among the rarer O-linked glycan modifications, enzyme contributing to O-mannosylation of cadherin superfamily (46), particularly TMTC1 was highly expressed in vascular endothelial cells (Supplementary Figure S5A). Not much is known about these pathways, but the measured high gene expression warrants additional investigation regarding its biological function. Among the enzymes involved in nucleotide biosynthesis, we noted high levels of UGDH (UDPGlcA_synth) and UXS1 (UDPXyl_synth), which are involved in the biosynthesis of starting materials that contribute to GAG biosynthesis (Supplementary Figure S5B). Enzymes in the biosynthesis of other nucleotide-sugar donors were also present in all cells, albeit at lower levels. Finally, several enzymes involved in lysosomal targeting and trafficking (lyso_target) of glycosidases were also highly expressed (Supplementary Table S4).

Among the enzymes mediating glycan extension, high expressions were noted in pathways involved in the biosynthesis of Type-II lactosamine (Galβ1–4GlcNAcβ) and Type-III lactosamine (Galβ1–3GalNAcβ) chains across all cell and tissue-types, compared to Type-I lactosamine (Galβ1–3GlcNAcβ) chains. This is generally consistent with current biological knowledge related to the high abundance of Type-II lactosamine chains on N-/O-linked glycans and GSLs. Interestingly, the abundance of I-branching enzymes GCNT2 and 3 were restricted to specific tissue types supporting the emerging notion that such GlcNAcβ1–6 branching may have important biological functions for example in regulating cell growth and survival (47). Additional extension glycopathways that were highly expressed were involved in the N-glycosylation processing enzymes that are expressed in the endoplasmic reticulum (Ngly_trim) and the keratan sulfate extension enzymes (KS_poly). The epithelial cells of the liver, which is a major source of heparan sulfate biosynthesis, also had expression of HS extension genes at high levels.

The expression levels of chain terminating enzymes were often low and heterogenous, across cell types. The only exception to this were the enzymes involved in protein tyrosine sulfonation (TyrP-synth), which were uniformly expressed at higher levels than other terminal glycopathways in all cells and tissues. This is consistent with the ubiquitous nature of this modification. Genes related to ABO antigen had the highest expression in epithelial cells (see Supplementary Table S4) as these cells are a major source of blood group antigens. Finally,

among the lowly expressed glycopathways, SDA antigen (Sda) biosynthetic enzymes were restricted to epithelial and immune cells while the enzymes forming sialyl Lewis-X epitope were dominant in T cells, monocytes and neutrophil populations based on expression of the enzyme FUT7.

The logFCs for glycopathway expression are generally moderate, ranging from –0.81 to 1.37 (see 10.5281/zenodo.14177056 for result of DE analysis), suggesting the prevalence of similar pathways in different cell types. Among the pathways, core and extension groups display greater logFC magnitudes than those in the terminal group suggesting heightened tissue-to-tissue variability. Similar to the terminal group, the core subclass, together with the NS and degradation groups, exhibits only modest logFCs across different tissues. Upon examining individual tissues, cells in the eye, heart and lung typically demonstrated higher overall glycopathway expressions when compared to other cells. In contrast, cells in the kidney and liver display relatively lower overall expressions. Interestingly, cells originating from tissues that are related to each other, such as large and small intestines, present similar expression patterns as shown by hierarchical clustering in Figure 2 and Supplementary Figure S4. Overall, several observations from comparing glycopathway expressions across tissues and cell types are consistent with literature and our analyses also suggest additional hypotheses that require experimental validation. To facilitate such comprehensive exploration of glycopathways, we have developed a web tool called glycoCARTA that is accessible from virtualglycome.org.

While DE analysis focuses on differences in mean gene expression, we complemented this with enrichment-depletion (ED) analysis. This analysis determines if a specific tissue is disproportionately enriched or depleted of cells that are expressing a given glycopathway—termed as 'expressing cells'—when compared to the overall proportion of expressing cells in the entire TS dataset. Given that a number of glycopathways demonstrated low single-cell expressions, the fraction of expressing cells has previously been proposed as a better metric for gauging expression within a cell population (40). To this end, for every combination of glycopathway and tissue, we constructed a contingency table showing the count distribution of expressing/non-expressing cells for the glycopathway and of cells from/outside of the tissue (see Figure 3A and 'Materials and methods' section). Utilizing this table, we evaluated the log2 odds ratio (logOR). A logOR > or <0 suggests that the particular tissue contains more or fewer expressing cells than expected based on the overall proportion of expressing cells in the TS dataset. In effect, a positive (negative) logOR, shown as a red (blue) colored dot in Figure 3, indicates enrichment (depletion) of cells expressing the designated glycopathway in that tissue. In the example shown in Figure 3A, there was enrichment of expressing cells for the GSL diversify pathway in the bladder, trachea and eye, i.e., these tissues have higher fractions of cells expressing the genes in this pathway than expected based on cells in the TS dataset. This observation is in good agreement with the DE analysis in Figure 2A, where the same three tissues (bladder, trachea and eye) have the highest mean expressions. Overall, the results of ED analysis as shown in Figure 3B resonate well with the DE analysis findings. Indeed, the logORs have a strong positive correlation with the logFC (Pearson's correlation, $\rho = 0.61$). With the exception of the GalCer pathway in the kidney, pathways with positive (negative) logORs generally had positive (negative) log2FCs.

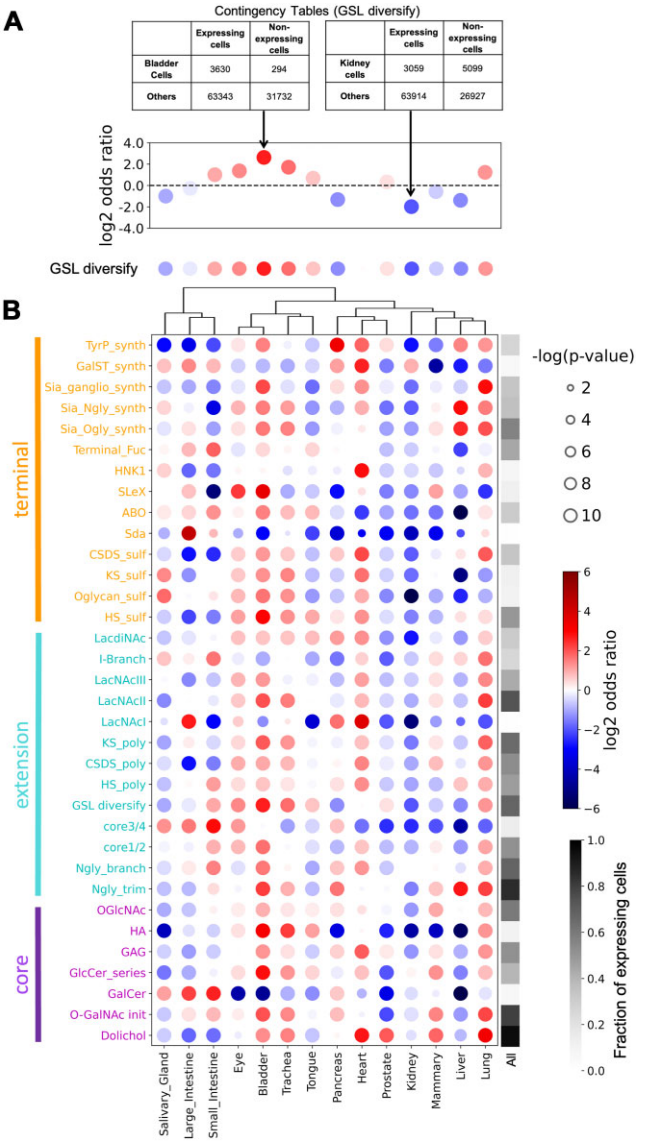

Figure 3. ED analysis of glycopathway in epithelial cells. (**A**) GSLs diversify pathway is used to illustrate calculation scheme. ED analysis was performed by constructing the contingency table. The size of the dot represents the *P*-value of the Fisher's exact test for significance, while the face color gives the sign of the log-odds ratio (logOR, blue: negative logFC and red: positive logFC). A negative logOR represents a depletion, while a positive logOR represents an enrichment of glycopathway in a tissue with respect to all other tissue with the same pathway. (**B**) ED of glycopathways in core, extension and terminal groups for epithelial cells. The last column presents the fraction of expressing cells for each glycopathway among all epithelial cells using grayscale heatmap. This allows evaluation of how prevalent a given pathway is in epithelial cells compared to other pathways.

In summary, the results of both DE and ED analysis highlight a moderate variability in glycopathway expressions across tissues, especially in the core and extension groups. Conversely, the terminal, core subclass, nucleotide sugar and degradation manifest higher variation in single-cell expression across different tissues. Additional wet-lab studies are warranted to determine how the variation in gene expression across tissue relate to tissue-specific glycan structure patterns.

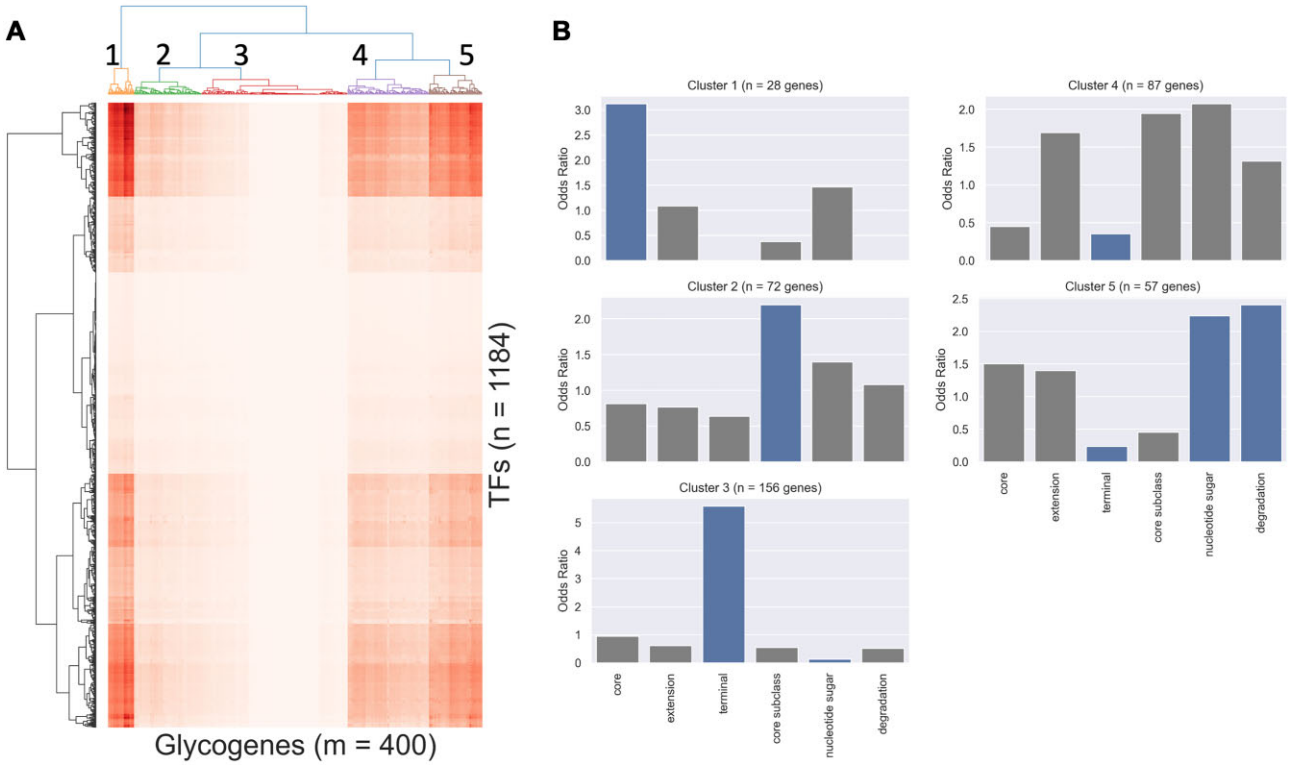

Figure 4. Glycosylation transcription factor analysis. (**A**) Hierarchical clustering of glycogenes based on the log1p of MI scores of TF–glycogene (i.e., (log(1 + MI)). The numbers indicate cluster labels. (**B**) Enrichment analysis of glycogenes in each cluster for glycopathway classes. Blue bars indicate those with Benjamini–Hochberg adjusted *P*-value < 0.1 (two-sided Fisher's exact test).

Transcriptional factors regulating glycosylation extracted from single-cell RNA-seq

Literature reported experimentally validated TFs regulating glycogenes are few (5). To address this gap, we leveraged single cell transcriptomics data in the TS and the TF–gene binding interaction data in TFLink database (37). Our strategy involved evaluating MI of single-cell gene expression between every possible pair of TF and glycogene. Here, MI gives a measure of how much the uncertainty in the expression of a glycogene is reduced given the corresponding expression data for a TF. Applying this strategy, the MI yielded an accurate prediction for TF–glycogene interactions, achieving an AUPRC of 0.447, when compared to TF-binding interactions sourced from the TFLinks (37). This accuracy outperformed both the pairwise Pearson's correlation (AUPRC = 0.367) and the Spearman's rank correlation (AUPRC = 0.396). All of the above interaction scores, MI, Pearson's correlation and Spearman's rank correlation, surpass the performance of a random predictor (AUPRC = 0.318, $P < 10^{-5}$). This outcome suggests that single-cell gene expression data may be used to infer TF–glycogene interactions.

Our single-cell TF–glycogene evaluation facilitates the identification of RMs of glycosylation. In this context, a RM refers to a set of glycogenes whose transcription is controlled by a shared regulatory program (48). To this end, we performed a hierarchical clustering of glycogenes using their log1p of MI scores with 1184 TFs (i.e., log(1 + MI)). The clustering reveals five RMs, as depicted in Figure 4A (see Supplementary Table S5 for glycogene membership in clusters). Subsequent analysis using Fisher's exact test linked each cluster with specific glycopathway classes (two-sided Fisher's exact test, Benjamini–Hochberg adjusted *P*-value < 0.1; see 'Materials and methods' section). The odds ratios presented in Figure 4B (see also Supplementary Figure S8) indicate that different RMs are associated with distinct classes of glycopathways, implying a shared transcriptional regulatory program among glycogenes from the same pathway class. We also curated a ranked list of TFs for each RM (see Supplementary Table S6), providing insights into potential regulatory factors. To explore the TF–glycogene analysis more fully, we developed a web-tool called glycoTF that is available at virtualglycome.org.

The first RM (Cluster 1) is strongly associated with the synthesis of Core glycan structures. The glycogenes in this cluster encode enzymes involved in the formation of the oligosaccharyltransferase (OST) complex (i.e., dolichol pathway): DAD1, DDOST, RPN1, RPN2 and STT3B; enzymes involved in initial processing of N-glycans: PRKCSH and GANAB; glycoprotein folding chaperones: CANX, CALR, ERLEC1, HSP90B1, HSPA5, IGF2R, LMAN2, OS9 and SE1L; O-GlcNAc biosynthesis enzymes: OGT and OGA; and other high abundance genes: B4GALT1, GPI, M6PR and MGAT1. The second RM (Cluster 2) is connected to a multitude of processes involves in the core subclass (see also Supplementary Figure S8). Specifically, glycogenes in Cluster 2 are involved in a number of glycopathways responsible for the biosynthesis of GAG and lipid-linked oligosaccharides, and the O-linked glycan post-translation modification such as POFUT and POMT genes. A number of transporter genes also belong to Cluster 2.

The third RM (Cluster 3) is significantly enriched for glycogenes involved in the terminal class. More specifically, cluster 3 includes UDP-Glucuronosyltransferase family genes that are involved in the glucuronidation that enable drug metabolism

and also metabolism of pollutants, bilirubin, androgens, estrogens, mineralocorticoids, glucocorticoids, fatty acid derivatives, retinoids and bile acids. This cluster also comprises a number of sulfotransferases that modify both GAGs and glycoproteins; a majority of genes (8 out of 11) that participate in the terminal protein fucosylation; and several members of the sialyltransferase family (10 out of 20). Importantly, the Terminal class is under-represented in all clusters besides Cluster 3 (OR < 1), suggesting that these processes are under a distinct transcriptional regulatory program than the others.

The fourth RM (Cluster 4) is intermixed with glycogenes from the NS metabolism, core subclass and extension groups, but none of these classes crossed the statistical significance cutoff (Benjamini–Hochberg adjusted *P*-value < 0.1). Important genes involved in the synthesis of nucleotide sugars, such as GALT, GALK2, GALE, PGM1-3 and GMPPA/B, belong to this RM. Another prominent feature of Cluster 4 is the presence of genes involved in the initiation of heparan sulfate and chondroitin sulfate biosynthesis including B4GALT7, FAM20B, B3GALT6, CHPF, CHPF2, EXT1 and EXT2. This cluster also includes genes involved in GPI anchor biosynthesis (PIGC, PIGG, PIGH, PIGK, PIGN, PIGS and PIGX) and in the modification of N-glycosylation, specifically in terminal sialylation and core-fucosylation (ST6Gal1 and FUT8). The reason why a single cluster of TFs would regulate a diverse group of pathways remains to be studied in literature.

The last RM (Cluster 5) is strongly linked to NS metabolism and degradation classes. Genes involved in NS biosynthesis in this cluster comprise CMAS, DPM1, DPM2, GALK1, GFPT1, GFUS, GNPDA1, PAPSS, UAP1 and UGDH. The cluster also includes a set of genes involved in the degradation processes, such as CTBS, CEMIP2, GLB1, GNS, GUSB, HEXA, HEXB, HGSNAT, IDS, NEU1 and FUCA2. Except for collagen degradation, all glycopathways in the degradation class are over-represented in this RM (OR > 1, see Supplementary Figure S8).

Discussion

A key contribution from this study is the description of the broad landscape of glycoEnzymes and glycopathways in normal human cell and tissue types. The findings are in broad agreement with the recent work by Joshi *et al.* (20) that showed that core enzymes are more ubiquitously expressed among cells than enzymes that modify terminal glycan residues which cater to more specialized functions. However, it is important to note differences in the study design as the glycogene set used in this work is considerably larger (224 in Joshi *et al.* versus 400 in this work), owing to the inclusion of glycosidases, transporters and other regulators of carbohydrate biosynthesis. In addition, the focus on glycopathways and well-defined ontologies represents a step away from the previous approach. Importantly, we also analyze single-cell gene expression data directly without pseudobulking. The use of DE and ED analysis, as opposed to using inter-quartile distances to judge the importance of specific glycogenes, is another difference. Finally, our study presents the first detailed analysis of TF–glycogene relations using single-cell gene expression data. This reveals the possibility that distinct TF RMs control various aspects of mammalian glycosylation.

Another contribution of our work is the online webtools for glycosciences, namely glycoCARTA for exploring glycogene and glycopathway expressions at single cell level and glycoTF for candidate transcriptional factors of glycosylation. Presently available online tools for single-cell gene expression data such as the CellXGene Discover (49) allow broad exploratory investigation of human transcriptome. Meanwhile, glycoCARTA enables a more targeted and in-depth examination of glycogenes and glycopathways in human, including comparisons of their expressions in different cell types and tissues. Further, in comparison with glycosylation-focused webtool Glycopacity by Joshi *et al.* that includes mainly glycosyltransferases (20), glycoCARTA covers a larger set of glycoEnzymes, as described above. We are not aware of any online resources for TFs of glycosylation.

While our study presents a broad analysis of human glycosylation pathways, it is not without limitations. At the sequencing depth employed in the TS, out of the anticipated ∼220 glycogenes expressed human cells, less than one fifth are detected in a typical cell in this dataset. While this low detection is comparable with other PC genes, such high data sparsity impedes data analysis such as single-cell clustering. The underlying TS data also do not include key organs like the brain which have distinct glycosylation profiles compared to other organs. Further data analysis is thus required to integrate the findings of this work with brain initiatives and related activities (50). In addition, while our analysis reveals the nuances of gene expression patterns related to glycosylation, additional wet-lab studies are needed to explore the causal mechanisms and also to extrapolate these findings to glycan structures on the individual cell types and related functional outcomes. Such an endeavor requires the development of novel technologies to measure glycoenzyme activity in greater depth at single cell level and parallel development of glycomics analysis. Despite limited coverage, recent advances in simultaneous single-cell sequencing of RNA and lectin binding (e.g. scGR-seq (23,51) and Sugar-Seq (21)) and the integrated analysis of the resulting data (22,52), have shown promise in employing single-cell analysis to elucidate the regulation of glycosylation (transcription, translation and biosynthetic reactions) and their impact on cell function. Thus, at this time, the observed variations in gene expression described in this manuscript only describe a portion of the factors affecting cellular variations in the glycome. Direct validation, such as through glycomics analysis at single-cell level and CRISPR (short for Clustered Regularly Interspaced Short Palindromic Repeats) technology based molecular screens are paramount in solidifying our inferences and bridging the gap between gene expression and functional glycan structures on proteins.

Data availability

The data underlying this article are available in Figshare at https://doi.org/10.6084/m9.figshare.14267219. The source codes are available in 10.5281/zenodo.14177056. Interactive webtools glycoCARTA and glycoTF are available at http://www.virtualglycome.org.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

National Institutes of Health [3R01HL103411-10S1, P01 HL151333]; SUNY Multidisciplinary Small Team Award [201047.2 to S.V. (in part)]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement

None declared.

References

1. Neelamegham,S. and Mahal,L.K. (2016) Multi-level regulation of cellular glycosylation: from genes to transcript to enzyme to structure. *Curr. Opin. Struct. Biol.*, **40**, 145–152.
2. Varki,A. (2017) Biological roles of glycans. *Glycobiology*, **27**, 3–49.
3. Reily,C., Stewart,T.J., Renfrow,M.B. and Novak,J. (2019) Glycosylation in health and disease. *Nat. Rev. Nephrol.*, **15**, 346–366.
4. Schjoldager,K.T., Narimatsu,Y., Joshi,H.J. and Clausen,H. (2020) Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.*, **21**, 729–749.
5. Groth,T., Diehl,A.D., Gunawan,R. and Neelamegham,S. (2022) GlycoEnzOnto: a GlycoEnzyme pathway and molecular function ontology. *Bioinformatics*, **38**, 5413–5420.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
7. Yamada,I., Shiota,M., Shinmachi,D., Ono,T., Tsuchiya,S., Hosoda,M., Fujita,A., Aoki,N.P., Watanabe,Y., Fujita,N., *et al.* (2020) The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nat. Methods*, **17**, 649–650.
8. Neelamegham,S. and Liu,G. (2011) Systems glycobiology: biochemical reaction networks regulating glycan structure and function. *Glycobiology*, **21**, 1541–1553.
9. Kellman,B.P. and Lewis,N.E. (2021) Big-data glycomics: tools to connect glycan biosynthesis to extracellular communication. *Trends Biochem. Sci*, **46**, 284–300.
10. Bennun,S.V., Hizal,D.B., Heffner,K., Can,O., Zhang,H. and Betenbaugh,M.J. (2016) Systems glycobiology: integrating glycogenomics, glycoproteomics, glycomics, and other 'Omics data sets to characterize cellular glycosylation processes. *J. Mol. Biol.*, **428**, 3337–3352.
11. Liu,G. and Neelamegham,S. (2015) Integration of systems glycobiology with bioinformatics toolboxes, glycoinformatics resources, and glycoproteomics data. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **7**, 163–181.
12. Huang,Y.F., Aoki,K., Akase,S., Ishihara,M., Liu,Y.S., Yang,G., Kizuka,Y., Mizumoto,S., Tiemeyer,M., Gao,X.D., *et al.* (2021) Global mapping of glycosylation pathways in human-derived cells. *Dev. Cell*, **56**, 1195–1209.
13. Farrell,J.A., Wang,Y., Riesenfeld,S.J., Shekhar,K., Regev,A. and Schier,A.F. (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, **360**, eaar3131.
14. Griffiths,J.A., Scialdone,A. and Marioni,J.C. (2018) Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.*, **14**, e8046.
15. Ginhoux,F., Yalin,A., Dutertre,C.A. and Amit,I. (2022) Single-cell immunology: past, present, and future. *Immunity*, **55**, 393–404.
16. Tian,Y., Carpp,L.N., Miller,H.E.R., Zager,M., Newell,E.W. and Gottardo,R. (2022) Single-cell immunology of SARS-CoV-2 infection. *Nat. Biotechnol.*, **40**, 30–41.
17. Perez,K., Ciotlos,S., McGirr,J., Limbad,C., Doi,R., Nederveen,J.P., Nilsson,M.I., Winer,D.A., Evans,W., Tarnopolsky,M., *et al.* (2022) Single nuclei profiling identifies cell specific markers of skeletal muscle aging, frailty, and senescence. *Aging (Albany NY)*, **14**, 9393–9422.
18. Tabula Muris,C. (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, **583**, 590–595.
19. Tabula Sapiens,C., Jones,R.C., Karkanias,J., Krasnow,M.A., Pisco,A.O., Quake,S.R., Salzman,J., Yosef,N., Bulthaup,B., Brown,P., *et al.* (2022) The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*, **376**, eabl4896.
20. Dworkin,L.A., Clausen,H. and Joshi,H.J. (2022) Applying transcriptomics to studyglycosylation at the cell type level. *iScience*, **25**, 104419.
21. Kearney,C.J., Vervoort,S.J., Ramsbottom,K.M., Todorovski,I., Lelliott,E.J., Zethoven,M., Pijpers,L., Martin,B.P., Semple,T., Martelotto,L., *et al.* (2021) SUGAR-seq enables simultaneous detection of glycans, epitopes, and the transcriptome in single cells. *Sci. Adv.*, **7**, eabe3610.
22. Minoshima,F., Ozaki,H., Odaka,H. and Tateno,H. (2021) Integrated analysis of glycan and RNA in single cells. *iScience*, **24**, 102882.
23. Odaka,H., Ozaki,H. and Tateno,H. (2022) scGR-seq: integrated analysis of glycan and RNA in single cells. *STAR Protoc.*, **3**, 101179.
24. Jiang,R., Sun,T., Song,D. and Li,J.J. (2022) Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.*, **23**, 31.
25. Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
26. Nairn,A.V., York,W.S., Harris,K., Hall,E.M., Pierce,J.M. and Moremen,K.W. (2008) Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes. *J. Biol. Chem.*, **283**, 17298–17313.
27. Groth,T., Gunawan,R. and Neelamegham,S. (2021) A systems-based framework to computationally describe putative transcription factors and signaling pathways regulating glycan biosynthesis. *Beilstein J. Org. Chem.*, **17**, 1712–1724.
28. Tabula Sapiens Consortium and Pisco,A. (2021) *Tabula Sapiens single-cell dataset*. https://doi.org/10.6084/m9.figshare.14267219.v5.
29. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
30. Yang,S., Corbett,S.E., Koga,Y., Wang,Z., Johnson,W.E., Yajima,M. and Campbell,J.D. (2020) Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.*, **21**, 57.
31. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M., *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
32. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
33. Gayoso,A., Lopez,R., Xing,G., Boyeau,P., Valiollah Pour Amiri,V., Hong,J., Wu,K., Jayasuriya,M., Mehlman,E., Langevin,M., *et al.* (2022) A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, **40**, 163–166.
34. McInnes,L., Healy,J., Saul,N. and Großberger,L. (2018) UMAP: uniform Manifold approximation and projection. *J. Open Source Software*, **3**, 861.
35. Miller,R.G. (1981) *Simultaneous Statistical Inference*. 2d edn., Springer-Verlag, New York.
36. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,,W.M. 3rd, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
37. Liska,O., Bohar,B., Hidas,A., Korcsmaros,T., Papp,B., Fazekas,D. and Ari,E. (2022) TFLink: an integrated gateway to access transcription factor-target gene interactions for multiple species. *Database (Oxford)*, **2022**, baac083.

38. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn Res.*, **12**, 2825–2830.
39. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
40. Booeshaghi,A.S. and Pachter,L. (2021) Normalization of single-cell RNA-seq counts by log(x + 1)dagger or log(1 + x)dagger. *Bioinformatics*, **37**, 2223–2224.
41. Chari,T. and Pachter,L. (2023) The specious art of single-cell genomics. *PLoS Comput. Biol.*, **19**, e1011288.
42. Wang,S., Sontag,E.D. and Lauffenburger,D.A. (2023) What cannot be seen correctly in 2D visualizations of single-cell 'omics data? *Cell Syst.*, **14**, 723–731.
43. Hart,G.W., Slawson,C., Ramirez-Correa,G. and Lagerlof,O. (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu. Rev. Biochem.*, **80**, 825–858.
44. Yang,J.M., Byrd,J.C., Siddiki,B.B., Chung,Y.S., Okuno,M., Sowa,M., Kim,Y.S., Matta,K.L. and Brockhausen,I. (1994) Alterations of O-glycan biosynthesis in human colon cancer tissues. *Glycobiology*, **4**, 873–884.
45. Brockhausen,I. (2006) Mucin-type O-glycans in human colon and breast cancer: glycodynamics and functions. *EMBO Rep.*, **7**, 599–604.
46. Larsen,I.S.B., Narimatsu,Y., Clausen,H., Joshi,H.J. and Halim,A. (2019) Multiple distinct O-mannosylation pathways in eukaryotes. *Curr. Opin. Struct. Biol.*, **56**, 171–178.
47. Colosimo,C., Rossi,P., Elia,M., Bentivoglio,A.R., Altavista,M.C. and Albanese,A. (1991) Transient alternating hemichorea as presenting sign of progressive supranuclear palsy. *Ital. J. Neurol. Sci.*, **12**, 99–101.
48. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.
49. CZI Cell Science Program, Abdulla,S., Aevermann,B., Assis,P., Badajoz,S., Bell,S.M., Bezzi,E., Cakir,B., Chaffer,J., Chambers,S., *et al.* (2024) CZ CELL×GENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.*, gkae1142.
50. Ament,S.A., Adkins,R.S., Carter,R., Chrysostomou,E., Colantuoni,C., Crabtree,J., Creasy,H.H., Degatano,K., Felix,V., Gandt,P., *et al.* (2023) The Neuroscience Multi-Omic Archive: a BRAIN Initiative resource for single-cell transcriptomic and epigenomic data from the mammalian brain. *Nucleic Acids Res.*, **51**, D1075–D1085.
51. Keisham,S., Saito,S., Kowashi,S. and Tateno,H. (2024) Droplet-based glycan and RNA sequencing for profiling the distinct cellular glyco-states in single cells. *Small Methods*, **8**, e2301338.
52. Qin,R., Mahal,L.K. and Bojar,D. (2022) Deep learning explains the biology of branched glycans from single-cell sequencing data. *iScience*, **25**, 105163.

Received: July 18, 2024. **Revised:** November 6, 2024. **Editorial Decision:** November 11, 2024. **Accepted:** November 21, 2024